UNIVERSITY OF TURKU

ALGORITHMIC FOUNDATIONS FOR GENERALIZABLE ARTIFICIAL INTELLIGENCE MODELS: A MULTI DOMAIN STUDY

Jatin K. Chaudhary

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. F62 | TECHNICA – INFORMATICA | TURKU 2025

University of Turku

Faculty of Technology
Department of Computing
Computer Science
The Doctoral Programme in Technology

Supervised by

Professor Jukka Heikkonen
Research Director
Department of Computing
University of Turku

Dr. Rajeev Kanth
Supervisor
Department of Computing
University of Turku

Dr. Harri Merisaari
Supervisor
Department of Diagnostic Radiology
University of Turku

Reviewed by

Prof. Andrej Škraba
Full Professor
Department of Informatics
Faculty of Organizational Sciences
University of Maribor
Slovenia

Prof. Durga Prasad Mohapatra
Full Professor
Department of Computer Science & Engineering
National Institute of Technology
India

Opponent

Prof. Pekka Toivanen
Full Professor
School of Computing
Faculty of Science
University of Eastern Finland

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-952-02-0259-0 (PRINT)
ISBN 978-952-02-0260-6 (PDF)
ISSN 2736-9390 (PRINT)
ISSN 2736-9684 (ONLINE)
Painosalama Oy, Turku, Finland

Om Namah Shivaya

My dedication goes to my family.

Abstract

The increasing deployment of artificial intelligence (AI) models in high-stakes domains such as medical diagnostics and renewable energy has exposed persistent limitations in their generalizability and reproducibility across varying data sources and computational settings. AI models often fail to maintain stable performance when transferred to new environments, largely due to distributional shifts, limited annotated data, and differences in hardware or acquisition protocols. This thesis addresses the critical challenge of building generalizable AI systems and provides a structured framework for evaluating and fine-tuning these models across diverse domains.

The core contribution of this dissertation is the development and analysis of foundational models that can be pretrained on heterogeneous datasets and subsequently adapted to new target domains using minimal site-specific data. These models are evaluated in two applied contexts: prostate cancer detection using MRI radiomics, and solar cell performance prediction using simulated photovoltaic datasets. In both domains, the research demonstrates that transfer learning, combined with carefully designed harmonization pipelines and explainability modules, leads to substantial improvements in cross-site adaptability and clinical or scientific relevance.

The theoretical component of the work explores the mathematical principles underlying generalizable learning. Drawing on concepts from Poisson's equations, continuity equations, and Lyapunov stability, the study proposes an exponential decay mechanism as a theoretically sound learning rate schedule for future implementation. This framework introduces the notion of equiconnectedness ensuring that the superlevel sets of the loss function remain connected under dynamic regularization thereby offering a foundation for convergence stability in overparameterized models. Although not applied in the practical domains presented, these mathematical insights provide direction for the principled design of future optimization strategies.

In the photovoltaic application, a Bayesian Regularized Neural Network is trained on over a million simulated configurations of silicon tandem multi-junction solar cells. The model accurately predicts performance metrics such as open-circuit voltage and fill factor, enabling data-driven optimization of solar devices under diverse environmental conditions. In the medical imaging application, a Vision Transformer-based model is pretrained on radiomic features extracted from multiparametric prostate MRI scans and fine-tuned on site-specific datasets. The model achieves an average AUC of 0.85, with improved results upon local adaptation, and demonstrates robust

performance across different scanner vendors and imaging protocols. Integration of SHAP and LIME further facilitates clinical interpretability.

In conclusion, this dissertation contributes to the development of generalizable AI systems by combining practical model-building strategies with rigorous mathematical theory. This dissertation advances the field of artificial intelligence by developing mathematically grounded optimization strategies that enhance neural network performance and stability. By successfully applying these strategies to complex, high-dimensional data landscapes in photovoltaics and medical diagnostics, the research demonstrates the broad applicability and impact of integrating rigorous mathematical theory with practical neural network design and optimization.

KEYWORDS: Generalizable Artificial Intelligence, Foundational Models, Transfer Learning, Vision Transformers (ViT), Radiomics, Prostate Cancer Detection, Bayesian Regularization, Neural Network Optimization, Fine-tuning Strategies, Harmonization Techniques, Explainable AI (XAI), Equiconnectedness, Lyapunov Stability, Exponential Decay (Theoretical), Photovoltaic Simulation, Domain Adaptation, Cross-site Validation, Multi-center MRI, Reproducibility in AI

Tiivistelmä

Tekoälymallien yleistyminen kriittisissä sovelluskohteissa, kuten lääketieteellisessä diagnostiikassa ja uusiutuvan energian optimoinnissa, on paljastanut merkittäviä puutteita mallien yleistettävyydessä ja tulosten toistettavuudessa erilaisissa aineistolähteissä ja laskennallisissa ympäristöissä. Usein tekoälymallien suorituskyky heikkenee, kun ne siiretään uusiin toimintaympäristöihin, pääosin jakauman muutosten, rajallisen annotoidun datan sekä laitteistojen tai tiedonkeruu- ja kuvantamisprotokollien erojen vuoksi. Tämä väitöskirja käsittelee tekoälyjärjestelmien yleistettävyyden keskeistä haastetta ja tarjoaa jäsennellyn viitekehyksen mallien arvioimiseksi ja hienosäätämiseksi monimuotoisissa sovelluskohteissa.

Väitöskirjan keskeinen kontribuutio on perustamallien kehittäminen ja analysointi. Nämä mallit voidaan esikouluttaa heterogeenisillä aineistoilla ja tämän jälkeen sovittaa uusiin kohdeympäristöihin minimaalisella, kohdekohtaisella datamäärällä. Mallien suorituskykyä arvioidaan kahdessa käytännön sovelluksessa: eturauhassyövän havaitsemisessa MRI-kuvantamisen radiomiikan avulla sekä aurinkokennojen suorituskyvyn ennustamisessa fotovoltaisten simuloitujen aineistojen pohjalta. Molemmissa sovelluskohteissa tutkimus osoittaa, että siirto-oppiminen yhdistettynä huolellisesti suunniteltuihin harmonisointiputkiin sekä tulosten tulkittavuuteen tähtääviin menetelmiin johtaa huomattaviin parannuksiin mallien sovellettavuudessa eri kohteissa sekä kliinisen tai tieteellisen merkityksen lisääntymiseen.

Työn teoreettinen osa tarkastelee yleistettävän oppimisen taustalla vaikuttavia matemaattisia periaatteita. Hyödyntäen Poissonin yhtälöihin, jatkuvuusyhtälöihin ja Lyapunovin stabiilisuuteen liittyviä käsitteitä, tutkimuksessa ehdotetaan eksponentiaalisen vaimennuksen mekanismia, joka toimii teoreettisesti perusteltuna oppimisnopeuden säätelystrategiana tulevaisuuden sovelluksissa.

Työn teoreettinen osa tarkastelee yleistettävän oppimisen taustalla vaikuttavia matemaattisia periaatteita. Hyödyntäen Poissonin yhtälöihin, jatkuvuusyhtälöihin ja Lyapunovin stabiilisuuteen liittyviä käsitteitä, tutkimuksessa ehdotetaan eksponentiaalisen vaimennuksen mekanismia, joka toimii teoreettisesti perusteltuna oppimisnopeuden säätelystrategiana tulevaisuuden sovelluksissa. Tässä yhteydessä esitellään ekvikonnektiviteetin käsite, joka varmistaa, että tappiofunktion tasoyläjoukot pysyvät yhtenäisinä dynaamisen regularisoinnin aikana. Tämä periaate tarjoaa perustan yliparametrisoitujen mallien konvergenssin stabiilisuudelle. Vaikka näitä teoreettisia havaintoja ei ole suoraan sovellettu väitöskirjan käytännön sovelluskohteisiin,

ne ohjaavat tulevaisuuden optimointistrategioiden suunnittelua vahvasti perustelluilla matemaattisilla näkökohdilla.

Fotovoltaisessa sovelluksessa Bayesian regularisoitu neuroverkko opetetaan yli miljoonalla simuloidulla piipohjaisen moniliitos-aurinkokennon konfiguraatiolla. Malli kykenee ennustamaan tarkasti suorituskykymittareita, kuten avoimen piirin jännitettä ja täyttökerrointa, mahdollistaen aurinkokennojen datalähtöisen optimoinnin erilaisissa ympäristöolosuhteissa. Lääketieteellisen kuvantamisen sovelluksessa visuaaliseen transformeriin (Vision Transformer, ViT) perustuva malli esikoulutetaan multiparametrisista eturauhasen MRI-kuvista johdetuilla radiomiikkaominaisuuksilla ja hienosäädetään kohdekohtaisilla aineistoilla. Malli saavuttaa keskimääräisen ROC-käyrän alaisen pinta-alan (AUC) 0,85, jonka tulokset paranevat edelleen paikallisen adaptoinnin myötä. Malli osoittaa robustia suorituskykyä eri laitevalmistajien ja kuvantamisprotokollien välillä. Lisäksi SHAP- ja LIME-menetelmien integrointi edistää kliinistä tulkittavuutta.

Johtopäätöksenä voidaan todeta, että väitöskirja edistää tekoälyjärjestelmien yleistettävyyttä yhdistämällä käytännönläheisiä mallinnusstrategioita ja vahvaa matemaattista teoriaa. Tämä tutkimus vie tekoälyn alaa eteenpäin kehittämällä matemaattisesti perusteltuja optimointistrategioita, jotka parantavat neuroverkkojen suorituskykyä ja vakautta. Soveltamalla onnistuneesti näitä strategioita monimutkaisiin ja korkealuotteisiin data-aineistoihin fotovoltaikan ja lääketieteellisen diagnostiikan aloilla, tutkimus osoittaa, miten laajasti sovellettavia ja vaikuttavia ovat matemaattisen teorian ja käytännöllisen neuroverkkosuunnittelun sekä optimoinnin yhdistämisen tulokset.

Acknowledgements

As I pen the final words of my doctoral journey, my heart swells with gratitude. This thesis is not just the product of experiments, data, or long nights of coding, it is the product of people, moments, and unwavering belief from those around me. Every chapter written, every milestone achieved, has been made possible through the shared strength, guidance, and love of many.

First and foremost, I extend my deepest thanks to my supervisors, Prof. Jukka Heikkonen and Dr. Rajeev Kanth. You both believed in me during the earliest and most uncertain phase of my career, when all I had to offer was curiosity, determination and persistence. Jukka, our coffee discussions were more than academic check-ins; they became my moments of certainty in an otherwise unpredictable research world. Rajeev, your kindness and the warmth of meals at your home gave me a sense of belonging, far away from mine. Thank you both for trusting in my potential, and for holding space for my growth even in times when I couldn't see it myself.

To Dr. Harri Merisaari, who joined my journey when I was at a crossroads, thank you for helping me rediscover direction when the path felt obscured. The hikes, and carols we took together weren't just escapes from academic stress; they were life lessons. You taught me how to endure, how to smile through the pain, and how happiness is often found in the simplest of moments.

To my beloved teachers, Late Zwala Dev, Mr. Sunny Sah, and Prof. Vipul Kheraj, thank you for believing in a dream that was not solely mine but ours together. Late Zwala, you saw the curious learner hidden within me, at a time when I disliked studying and never imagined enjoying science. Your gentle encouragement forever changed my path, and though you are no longer here, your influence remains deeply alive within me today. Mr. Sunny Sah, your passion for scientific exploration ignited my love for experiments and genuine discovery. You taught me not just how to dream big, but how to face setbacks without losing heart. Your lessons continue to guide me, especially in moments of struggle. I still carry your advice every time things get tough: don't lower the aim, just learn and try again. Prof. Kheraj, thank you for trusting a young and inexperienced student with your lab and precious resources. You gave me the first opportunity to pursue my curiosity and introduced me to the realities of research. Your support marked the true beginning of my journey as a researcher. Today, as I complete my PhD thesis, I carry your dreams with mine, deeply grateful and inspired by each of you.

To my parents, whose dreams became mine, and whose strength became my backbone, thank you for gifting me a vision of becoming a scientist, long before I saw it myself. Even during the crossovers of life when you were physically absent, your teachings and silent blessings stood by me like a lighthouse in the fog. To my sister, Miss. Jiya Chaudhary, thank you for being the ear to my endless frustrations, my emotional compass and the voice that always told me, "You've got this." For every stormy night I unloaded my resentments onto you, for every tiny win you celebrated as if it were your own, you have always been my safe haven.

To my colleagues, Mr. Adrian Borzyszkowski, Dr. Luca Zelioli, Mr. Dipak Nidhi, Mr. Javad Sheikh, Dr. Nitin Bayal and Dr. Dattatray Mongad, thank you for your camaraderie and kindness. Adrian, your ted talks, festival celebrations and beer meetups kept me sane all these years, Luca, your strength, and support meant more to me than I often expressed. Dipak, Nitin, and Dattatray our legendary WhatsApp group *Panchayat* wasn't just a chatroom; it was a lifeline. The laughter, the late night memes, and the unfiltered truths shared there helped me survive more than I can admit. I'll always cherish that space where we could be raw, real, and ridiculous all at once.

To Miss. Helen Pervez, how do I thank someone who became my family in a foreign land? From day one of this journey to the final day of defense, you were there. You've held my fears, dried my tears, and celebrated my every inch of progress with unshakeable loyalty. You've endured the daily refrain, *"I can't do it. I'm quitting and heading back to India."*, and never once let me give up. It was you who could change *"I am quiting"* to *"Lord shiva, I got a Jufo 3 publication"*. You saw me through every late night panic, and every self doubt. This thesis carries your spirit as much as mine. I owe this milestone to your belief in me when mine wavered. You are, and will always be, home.

To my constants, Mr. Ujjuwal Abhishek, Mr. Prashant Kamat, Dr. Sneha Gupta, Miss. Pragya Narayana Prasad, Mr. Swastik Bhattacharya, and Dr. Ananyo Bhattacharya, thank you for turning this long, winding academic journey into something profoundly human. Thank you for listening, for showing up, and for reminding me of the world outside academic deadlines. Ujjuwal, how do I even put this into words? You weren't just there from the beginning, you are the beginning. We didn't meet; we arrived together. From childhood mischief and muddy knees to now, as I teeter on the edge of becoming *"Dr. Me"*, you've been by my side. You've held my doubts, cheered my wins, and never let me forget who I was, even when I nearly did. You're not a friend, you're a living thread in the fabric of my life. Prashant, you've been on my side through the awkward years of school, the lost years, and to present times. You've been a silent guardian of my chaos, never loud, never absent. Our bons isn't just rare, it's grounding. You've been the kind of friend who doesn't need to be asked to stay; you just always have (with alcohol, obviously!). Sneha, you've been with me since the very seed of this dream was planted, back in undergrad when I first

whispered to the falling star that I wanted to become a researcher. You've seen this dream grow, take shape, bend under pressure, and stand tall again. To Pragya, thank you for starting our first research project together. You will always be my favourite lab mate. Swastik, my co-author and chaos-calmer, thank you for diving headfirst into every pit I managed to fall into, often in a more panicked state than I was. Your rescue missions have been uncountable, and heroic. Ananyo, your unwavering faith in real science and real scientists (including us!) gave me something to hold on to when despair crept in. You've kept the fire of curiosity alive. You six are not just the pillars of this PhD, you are the drivers of my life. Of every moment, every meltdown, every miracle. Thank you for being the people I could fall back on, and the ones I'll carry forward with me, always. Thank you for seeing me through.

And lastly, to Lord Shiva, the one whose presence I've carried every day. The one whose image rests on my wallpaper, to whom I bow each morning, and whose name I whispered in every moment of despair and joy. You made me. You drive me. You've given me strength beyond comprehension, and clarity when all else felt blurred. Every line in this thesis is a testament to your grace.

This chapter marks the end of a formal journey, but it also marks a beginning. And I carry with me not just the learnings, but every smile, every struggle, every blessing that made this possible.

Har Har Mahadev.

Jatin Chaudhary
Turku, Finland
May 2025

Table of Contents

Abbreviations

ADC	Apparent Diffusion Coefficient
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CDS	Clinical Decision Support
CDSS	Clinical Decision Support Systems
DWI	Difusion Weighted Imaging
FF	Fill Factor
GBR	Gradient Boosting Regressor
GLCM	Gray Level Co-occurrence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
IBSI	Image Biomarker Standardization Initiative
IG	Information Gain
MI	Mutual Information
MPP	Maximum Power Point
MRMR	Maximum Relevance Minimum Redundancy
MSE	Mean Squared Error
NLP	Natural Language Processing
PACS	Picture Archiving and Communication Systems
PV	Photo-Voltaics
QE	Quantum Efficiency
RBF	Radial Basis Function
RF	Random Forest
RFE	Recursive Feature Elimination
RFR	Random Forest Regressor
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RRMSE	Relative Root Mean Squared Error
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
SVR	Support Vector Regression

TGQO	Transcendental Genetic Quantum Optimization
XAI	Explainable AI

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Chaudhary, J., Jambor, I., Aronen, H., Ettala, O., Saunavaara, J., Boström, P., Heikkonen, J., Kanth, R. and Merisaari, H., Can Radiomics Based Models Survive Across MRI Scanners? Accepted for Publication in Springer Lecture Notes in Networks and Systems.

II Chaudhary, J., Nidhi, D., Heikkonen, J., Merisaari, H. and Kanth, R., 2024. Super Level Sets and Exponential Decay: A Synergistic Approach to Stable Neural Network Training. Journal of Artificial Intelligence Research (Accepted for Publication)

III Chaudhary, J.K., Liu, J., Skön, J.P., Chen, Y.W., Kanth, R.K. and Heikkonen, J., 2019. Optimization of silicon tandem solar cells using artificial neural networks. In Artificial Intelligence XXXVI: 39th SGAI International Conference on Artificial Intelligence, AI 2019, Cambridge, UK, December 17–19, 2019, Proceedings 39 (pp. 392-403). Springer International Publishing.

IV Chaudhary, J., Bhattacharya, S., Heikkonen, J. and Kanth, R., 2022, June. Prediction of Electron Band Gap of A_2XY_6 Perovskite Compounds using Machine Learning. In 2022 IEEE 49th Photovoltaics Specialists Conference (PVSC) (pp. 1173-1176). IEEE.

V Chaudhary, J., Jambor, I., Aronen, H., Ettala, O., Saunavaara, J., Boström, P., Heikkonen, J., Kanth, R. and Merisaari, H., Foundational AI and Radiomics: Improving Reproducibility in Clinical Decision Support Systems for prostate MR, Nature Cancer, 2025 (submitted).

1 Introduction

Artificial Intelligence (AI) has emerged as a transformative force across multiple scientific and industrial domains, reshaping how we approach diagnostics in healthcare, optimization in energy systems, and discovery in materials science. Fueled by advances in deep learning, AI systems have demonstrated exceptional performance in tasks that were once considered prohibitively complex [1; 2]. This rapid evolution has not only catalyzed academic exploration but has also brought AI technologies closer to real-world deployment in critical sectors. Yet, amid this progress, a foundational challenge persists: the inability of many AI models to maintain performance across diverse, unseen environments. While models may excel under controlled conditions or within narrowly defined datasets, their reliability often degrades when applied to new tasks, datasets, or institutional contexts. This problem is particularly pronounced in high stakes domains like clinical diagnostics, where performance variability can have significant consequences, and in scientific modeling, where AI driven predictions must generalize across varied material or environmental parameters [3; 4; 5]. The issue of generalizability is not merely a technical obstacle it represents a fundamental limitation in the current paradigm of model development. Existing literature often prioritizes accuracy within constrained benchmarks [6; 7], neglecting the broader goal of building adaptable, reproducible, and stable AI systems that can operate reliably across heterogeneous settings. This limitation hinders the practical adoption of AI in fields where data distributions shift frequently and retraining models for each new scenario is infeasible or prohibitively expensive.

The urgency to solve this problem has become more pronounced with the growing deployment of AI in real-world applications. As AI systems move from laboratory environments to clinical institutions, renewable energy labs, and national infrastructure, their success hinges not just on peak performance, but on robust, domain agnostic adaptability. This thesis is motivated by that urgent need to move beyond narrow optimization and contribute to the creation of foundational AI models that generalize effectively, are theoretically grounded, and can be trusted across domains.

Over the past decade, significant strides have been made in the development and deployment of AI systems, particularly those driven by deep learning architectures. These advancements have enabled high-performance models in medical imaging, natural language processing, renewable energy forecasting, and scientific simulations. For instance, deep convolutional neural networks (CNNs) and their derivatives

have formed the backbone of computer vision systems, excelling in domain-specific tasks with substantial labeled data [1; 2].

In the domain of medical diagnostics, especially radiomics-based analysis for prostate cancer, deep learning models have demonstrated high diagnostic accuracy when trained and tested on datasets from the same institution [3; 4]. Similarly, in energy science, machine learning techniques have shown considerable promise in optimizing photovoltaic cell configurations under simulation settings, particularly for silicon-based tandem architectures [5]. These successes underscore the power of AI to model complex, nonlinear relationships across high-dimensional spaces. Despite these advances, major limitations remain unresolved most notably the lack of generalizability. AI models that perform well in one context often exhibit sharp declines in accuracy when exposed to slightly different conditions or data sources. In radiomics, this manifests as poor reproducibility across imaging centers, stemming from scanner variations, acquisition protocols, and annotation biases [8]. In photovoltaic modeling, AI systems frequently fail to extend their predictive capabilities to new materials or configurations not seen during training [9]. These limitations underscore a deeper problem: current models are often overfit to narrow domains, optimizing performance within static boundaries rather than learning representations that are stable across shifts in data distribution. Moreover, while the adoption of transformer-based models such as Vision Transformers (ViTs) has sparked optimism due to their flexible attention mechanisms and ability to encode global context [10], their success is still highly dependent on curated training strategies, harmonized datasets, and extensive computational resources. Without careful design, even these models risk replicating the same failures in generalization that plagued earlier architectures. From a theoretical standpoint, model training continues to be guided largely by empirical heuristics, such as fixed learning rate schedules, without deep integration of stability principles. Although recent work has begun to explore the use of exponential decay for improving convergence and avoiding instability [11; 12], these methods are rarely grounded in rigorous mathematical frameworks that explain why they work, and when they fail.

While the field has achieved notable milestones, critical questions remain open: How can we build AI models that generalize reliably across domains? What theoretical tools can help us understand and guide the training of such systems? And how can we bridge the gap between high empirical performance and trustworthy, reproducible deployment?

While existing studies have made notable progress in optimizing AI models for specific domains, a fundamental challenge persists: the absence of robust, theoretically grounded frameworks for generalization across diverse environments. Current literature emphasizes empirical performance, often benchmarked on static or internally consistent datasets. This narrow focus has led to models that are brittle when confronted with unseen conditions, such as cross-institutional medical data or out-of-

distribution materials in solar energy optimization [8; 9]. A common thread in these failures is the lack of domain-agnostic architecture design and principled learning strategies. Most models are engineered with implicit assumptions about the training data distribution, and little effort is made to understand the underlying dynamics of training in non-stationary settings. Even in emerging approaches like transformer-based architectures, which offer improved flexibility [10], generalization remains fragile unless paired with extensive fine-tuning or data harmonization methods that are often impractical or costly in real-world deployments. Furthermore, although techniques such as exponential decay in learning rates have been heuristically applied to stabilize training [11], there is a lack of rigorous theoretical integration. Few studies connect these optimization strategies to mathematical tools like Lyapunov stability or probabilistic guarantees that can explain their behavior in dynamic training environments [12]. As a result, many improvements remain empirical and ad hoc, with limited understanding of their generalizability. This gap is particularly problematic in mission-critical applications such as clinical diagnostics, where reproducibility and interpretability are non-negotiable, and in engineering systems, where models must adapt to complex physical constraints and real-world variability. There is a pressing need for AI systems that not only perform well but are inherently designed for adaptability, reproducibility, and theoretical soundness. This thesis addresses a critical shortcoming in the field: the lack of unified AI frameworks that integrate generalizable architectures, empirically validated training protocols, and mathematically grounded optimization methods. In doing so, it aims to move beyond task-specific excellence and toward a principled foundation for cross-domain AI.

This thesis aims to address the above gaps and challenges by developing AI models that are generalizable, reproducible, and theoretically grounded across domains. The central objective is to move beyond narrow, domain-specific optimization and create foundational AI systems that perform robustly across diverse datasets and applications, particularly in healthcare diagnostics and energy optimization. The novelty of this work lies in its integrated approach that combines three critical pillars:

- Transformer-based Architectures: Building on the recent success of Vision Transformers (ViTs), this thesis leverages their capacity for attention-based feature learning to construct models that can generalize across heterogeneous data, such as multi-institutional medical images or simulation-driven solar cell datasets [10]. By explicitly modeling long-range dependencies and global context, these architectures offer a flexible backbone for domain transfer.

- Theoretically Informed Training Strategies: A key innovation in this research is the application of exponential decay as a foundational training mechanism rather than a heuristic. This decay is interpreted through the lens of Lyapunov

stability analysis and control theory, providing a principled framework to improve convergence, stability, and adaptability during optimization [11; 12].

- Domain-Aware Evaluation: Unlike many prior works that evaluate models in isolated settings, this thesis conducts cross-domain evaluations across radiomics-based clinical data and photovoltaic simulations. These domains differ significantly in structure and context, yet both demand models that can operate reliably under variable input conditions. This dual-application design serves as a rigorous testbed for the proposed generalization strategies.

Together, these contributions form a unified framework that tackles generalization not just from an algorithmic perspective but also through empirical validation and theoretical rigor. The thesis does not treat explainability, stability, and performance as isolated objectives but seeks to synthesize them into a cohesive modeling philosophy that supports real-world deployment of AI systems. In doing so, this work contributes to a growing body of research that aspires to make AI models not only more powerful but also more trustworthy, interpretable, and broadly applicable.

1.1 Research Questions

To operationalize the goals of this thesis and address the challenge of developing generalizable and theoretically grounded AI systems, the following research questions are posed. These questions guide the empirical investigations and theoretical developments presented in the subsequent chapters:

- What is the role of transformer-based architectures in enhancing the robustness and reproducibility of AI models when applied across heterogeneous datasets?

- In what ways can foundational AI models be fine-tuned with minimal domain-specific data while maintaining their performance across varying protocols and variations?

- To what extent can optimization techniques, such as Artificial Neural Networks (ANNs), predict efficient configurations in multi-junction solar cells using limited data?

- What are theoretical foundations which can be used for model training with high variation data?

- How can exponential decay mechanisms be integrated into the optimization process of deep learning models to improve their generalization across domains (e.g., medical imaging and photovoltaics)?

Each question corresponds to a focused investigation presented through individual research articles that are integrated within this thesis. Collectively, these questions aim to bridge the gap between empirical effectiveness and theoretical soundness, ensuring that the developed AI models are not only high-performing but also stable, reproducible, and applicable across a broad spectrum of real-world conditions.

1.2 Key Contributions

This thesis demonstrates that it is possible to design AI models that maintain high performance across domains by integrating architectural flexibility, training stability, and rigorous cross-domain evaluation. The contributions span both applied and theoretical domains and are supported by empirical evidence from two distinct application areas: medical imaging and solar energy optimization.

First, the thesis shows that transformer-based architectures specifically Vision Transformers can outperform traditional convolutional models in cross-institutional clinical settings when supported by proper feature selection and harmonization methods. Using radiomics features from multi-center prostate MRI datasets, the models were able to generalize well across imaging protocols and scanner types, addressing a major reproducibility challenge in medical AI.

Second, in the domain of energy science, the research demonstrates that Artificial Neural Networks trained with Bayesian regularization can reliably predict optimal configurations of silicon tandem solar cells. This result is significant because it offers a data-efficient alternative to traditional simulation-based optimization, thereby accelerating the design cycle for next-generation photovoltaic devices.

Third, the thesis provides a theoretical framework that elevates exponential learning rate decay from a practical heuristic to a principled mechanism. Through Lyapunov-based analysis, it is shown that decay schedules aligned with stability criteria can reduce training instability, prevent divergence, and improve generalization in models exposed to high data variability.

Together, these contributions suggest that building generalizable AI systems does not require trading off performance for flexibility. Instead, by unifying architectural design, optimization theory, and empirical validation, this thesis lays a foundation for developing AI models that are both adaptable and robust capable of performing reliably in complex, real-world scenarios without extensive domain-specific customization.

1.3 Organization of the Thesis

The thesis is structured into nine chapters, each building upon the previous to construct a cohesive narrative around generalizability in AI.

- **Chapter 1** is the introduction of the thesis. This chapter introduces the problem context, central research questions, and outlines the methodological and theoretical motivations.
- **Chapter 2** reviews the background literature across three domains: radiomics in medical imaging, solar energy optimization using AI, and theoretical aspects of training stability in neural networks.
- **Chapter 3** contains the datasets and data Preprocessing strategies used in this thesis.
- **Chapter 4** presents the detailed optimization strategy for neural networks. This chapter presents the theoritical contributions done by the candidate in the thesis. This chapter focused on theoretical AI, presenting proofs and experimental validation of the benefits of exponential decay in model training.
- **Chapter 5** includes the first empirical study on prostate cancer diagnosis using a transformer-based radiomics model trained across multicenter data.
- **Chapter 6** presents the second empirical study on optimization of silicon tandem solar cells using ANN models trained with Bayesian regularization and validated using SCAPS.
- **Chapter 7** presents the contribution of this thesis to the scientific fraternity. This chapter talks about different articles published by the candidate, his contribution in each of them, and the key results.
- **Chapter 8** discusses the results and outcomes of research undertaken towards this thesis.
- **Chapter 9** concludes the thesis by summarizing findings, discussing limitations, and proposing directions for future research, particularly in the design of scalable and generalizable foundational models.

2 Background

Statistical and probabilistic methods have profoundly shaped the theoretical and applied development of artificial intelligence (AI) models. In the context of this thesis, statistical modeling refers to the use of formal mathematical frameworks to understand and quantify relationships between variables, often under assumptions about data distributions, with a focus on interpretability and uncertainty estimation. Unlike purely descriptive statistical models, which summarize observed data, or traditional machine learning models, which emphasize predictive performance on unseen data, the statistical approaches employed here serve a dual role: they contribute to both model interpretability and predictive generalization. Probabilistic methods, particularly Bayesian inference, complement these models by enabling principled reasoning under uncertainty and allowing prior domain knowledge to be incorporated into the learning process. These frameworks have been especially influential in high-stakes domains such as medical diagnostics and renewable energy forecasting, where both accuracy and reliability are essential. Their integration into AI pipelines has enhanced the robustness, reproducibility, and adaptability of predictive systems across heterogeneous datasets and deployment environments [13; 14]. Throughout this thesis, statistical and probabilistic approaches have been central to both the theoretical underpinnings and practical implementations of foundational AI models.

2.1 Literature on Statistical and Probabilistic Approaches in AI

Statistical modeling in artificial intelligence refers to the use of mathematical constructs to represent the structure and relationships within data, enabling systems to make predictions, learn from observations, and manage uncertainty. These models are designed to infer generalizable patterns from data rather than rely on manually encoded rules, thus making AI systems more flexible and scalable across domains. For instance, rather than using fixed if-then conditions, statistical models such as those based on regression, Bayesian inference, or neural networks learn distributions and dependencies directly from training samples [13].

In contrast to traditional rule-based systems, which are limited in adaptability and require exhaustive knowledge engineering, statistical methods allow AI to adapt by capturing probabilistic associations in the data. This capability is particularly

relevant in dynamic and complex environments, such as medical diagnostics or solar energy forecasting, where inputs may vary significantly across time or contexts [5; 15]. In clinical applications, for example, statistical models have proven essential in dealing with heterogeneous MRI data across multiple vendors, allowing AI systems to perform consistently in prostate cancer detection regardless of imaging protocols [16].

Despite their strengths, statistical models face challenges such as overfitting, high computational demands, and the requirement for diverse datasets. These limitations can adversely affect real-world applications. For instance, in the context of prostate cancer diagnosis using MRI data, machine learning models trained on datasets obtained from a restricted subset of scanner vendors or imaging protocols often exhibit limited external validity. These models may perform adequately within the original training environment but fail to produce reliable predictions when applied to data from different scanners or institutions. Such lack of generalizability can result in diagnostic inaccuracies, thereby delaying appropriate clinical interventions and adversely affecting patient outcomes. Our foundational Vision Transformer (ViT)-based model addressed this by incorporating radiomic features from over 1,100 patients across varied institutions and scanner types, thus improving generalizability and clinical reproducibility [15].

Bayesian probability theory is a key pillar of statistical AI. It supports decision-making under uncertainty by updating beliefs as new data becomes available. This is particularly useful in domains like solar power forecasting and medical imaging, where uncertainty is intrinsic to the problem space [17; 18]. Bayesian Neural Networks (BNNs), an extension of this framework, integrate probabilistic reasoning directly into neural architectures, allowing models to produce not only predictions but also confidence intervals. In our photovoltaic study, BNNs trained using Bayesian regularization yielded more stable and interpretable current density estimates under varying environmental conditions compared to standard neural networks [5].

In healthcare, Bayesian modeling contributes to safer clinical AI. For instance, in our radiomics-based prostate cancer diagnostic system, we employed SHAP and LIME (two python based tools used for explainable AI) to interpret predictions at both global and patient-specific levels, aligning model behavior with clinical expectations. These explainable outputs are vital in contexts where model transparency directly affects clinical trust and adoption.

Throughout this thesis, statistical and probabilistic approaches have been central. In our radiomics work, statistical harmonization techniques such as quantile transformation and z-score normalization were used to reduce site-specific biases. We also employed mutual information and minimum redundancy maximum relevance (mRMR) for feature selection, ensuring that only the most informative and non-redundant features were retained. Similarly, in the photovoltaic study, an Artificial Neural Network trained with Bayesian regularization was used to optimize

silicon tandem solar cells. The model efficiently predicted optimal power outputs using fewer computational resources than conventional simulation methods [5].

In summary, statistical and probabilistic methods form the backbone of adaptive and trustworthy AI systems. Their role in this thesis is twofold: enabling reliable predictions across diverse datasets and providing mechanisms to quantify and interpret uncertainty. These contributions are critical in ensuring that AI models remain robust, interpretable, and applicable to real-world scenarios such as clinical diagnostics and energy optimization.

2.2 Generalization of a model, and its impact

Generalization is a fundamental property of artificial intelligence (AI) models, referring to their capacity to maintain predictive accuracy when applied to new, previously unseen data. This ability is critical for real-world deployment, where the data distribution encountered during inference often differs sometimes substantially from the training distribution. In this context, generalization is influenced not only by the model's structure but also by the relationship between the number of parameters and the size and variability of the training data. Models with too many parameters relative to the available data may overfit, capturing noise rather than signal, whereas overly simplistic models may underfit, failing to capture relevant patterns. To understand and improve generalization performance, this section examines key theoretical concepts including the bias-variance tradeoff, Occam's razor, and model complexity. The bias-variance tradeoff captures the tension between two sources of error in a predictive model: bias, which refers to systematic error due to overly simplistic assumptions, and variance, which reflects sensitivity to fluctuations in the training data. A model with high bias may fail to capture essential patterns (underfitting), while one with high variance may tailor itself too closely to the training data (overfitting). Achieving good generalization requires finding an appropriate balance between these two extremes. Occam's razor complements this view by favoring models that achieve predictive success with fewer assumptions or lower complexity, aligning with the principle that simpler models are more likely to generalize well. Together, these principles form a theoretical basis for designing models that are robust, interpretable, and suitable for deployment in dynamic, data-diverse environments [19].

The bias-variance tradeoff is essential for understanding generalization. It decomposes a model's prediction error into three components: bias (error from erroneous assumptions), variance (error from sensitivity to training data fluctuations), and irreducible error (inherent noise in the problem). Striking a balance between bias and variance is critical for achieving optimal performance [19]. An illustrative example of the bias-variance tradeoff is fitting polynomial curves of varying degrees to a dataset and analyzing the changes in training and validation errors. Occam's razor advocates for simplicity by favoring models with fewer parameters to avoid

overfitting, which can lead to better generalization [20]. In practice, regularization techniques such as L2 regularization, which penalizes model complexity by adding a term to the loss function, are commonly used to enforce simplicity and enhance generalization [21].

Neural networks, renowned for their capacity to capture complex patterns, present unique challenges concerning generalization and reproducibility. It is crucial to connect theoretical concepts to practical implementations within these models, as neural networks often involve sophisticated architectures and training regimes. Regularization techniques like dropout, where random neurons are ignored during training to prevent overfitting, are vital in improving the generalization capabilities of neural networks [22]. Similarly, data augmentation artificially enriches the training set by applying transformations such as rotation and scaling, thereby making models robust to input variations [23]. The reproducibility of neural networks is paramount, especially given the stochastic nature of their training processes, which depend significantly on initial conditions, datasets, and hyperparameters [parameter is learned by the model during training (e.g., weights in a neural network), whereas a hyperparameter is set before training and controls the learning process (e.g., learning rate, number of layers)]. Consistent reproducibility requires stringent standardization protocols, as emphasized in research on cross-vendor validation and robustness testing in medical imaging AI models [24; 25]. In medical imaging, the integration of multimodal features such as radiomic, morphometric, and intensity-based data types, rather than simply multiple scans from the same subject enhances generalizability, as seen in deep learning frameworks applied to prostate cancer detection using MRI [26]. These frameworks necessitate thorough cross-validation to ensure reliability across different medical datasets. To address generalization challenges, computational stress testing has been proposed to explore models' robustness and identify underspecification issues, thereby enhancing their suitability for diverse application domains [27]. In the broader context of AI applications in healthcare, particularly medical imaging, ensuring both generalization and reproducibility is imperative for the reliability and effectiveness of clinical decision-making. These two properties are tightly coupled with the concept of standardization, which refers to the consistent handling of data acquisition, preprocessing, feature extraction, model training, and evaluation protocols across different clinical settings. Standardization efforts aim to minimize variability introduced by differing MRI scanner vendors, imaging protocols, annotation methods, and even computational infrastructure. This includes harmonizing radiomic features through normalization techniques, using shared ontologies for labeling, and adhering to unified software pipelines for reproducible analysis. As highlighted by experts in the field, such consistency is essential for enabling AI models to be transferable across institutions, scalable to larger populations, and trustworthy for integration into routine clinical workflows [28]. Without standardization, even high-performing models risk failing in external validations,

thereby limiting their clinical utility.

2.3 Theoretical Framework

2.3.1 Partial Differential Equations

Differential equations play a pivotal role in modeling various aspects of artificial intelligence (AI), especially in the domains of neural networks and sophisticated machine learning algorithms. These mathematical constructs are essential for describing the dynamic changes within AI systems, facilitating a deeper understanding of how data is processed and information is transmitted across intricate network architectures. Poisson's equation, a fundamental partial differential equation (PDE) in electrical engineering and physics, finds significant applications in modern AI by modeling how electrical potentials are distributed in neural networks. The equation is expressed as:

$$\nabla^2 \phi = -\rho \tag{2.1}$$

Here, ϕ denotes the electrical potential, and ρ represents the charge density. Within the context of neural flows(here, neural flows refer to neural network models that learn continuous transformations of data using differential equations, typically parameterized by neural networks, allowing smooth and invertible mappings between input and output distributions), ϕ can be interpreted as the signal potential at any point in the neural network, while ρ signifies the density of the signal or information being processed. The utilization of Poisson's equation allows AI researchers to model and predict the behavior of neural signals within layers of a deep learning network. The distribution of these potentials critically influences the efficiency and accuracy of data processing and transmission, thereby optimizing algorithms that depend on the propagation of electrical signals, such as those used in training convolutional neural networks (Chaudhary, 2019).

Continuity equations are crucial for ensuring the conservation of certain quantities, such as mass or energy, within dynamic systems. In AI systems, particularly concerning neural flows, these equations are paramount for maintaining the integrity and continuity of information:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{2.2}$$

In this equation, ρ represents the information density, and $\mathbf{v}$ is the velocity vector field of the information flow across the network. Application of continuity equations ensures that during complex computations within AI models, no information is lost or artificially created, thus preserving the system's stability and reliability [29]. Maintaining the integrity of information within AI systems is critical, especially in

applications requiring high levels of accuracy and reliability such as medical imaging and autonomous driving. Differential equations like Poisson's and continuity equations provide a robust framework for understanding and designing systems that effectively manage the flow and preservation of information. Through precise modeling of neural flows, researchers can prevent the degradation of crucial data during processing, leading to more robust and generalizable AI models. The mathematical insights gained from these equations facilitate the optimization of neural network architectures, enhancing their performance and efficiency across various tasks and datasets [5].

2.3.2 Statistical Modelling

Statistical methods are integral to numerous Artificial Intelligence (AI) applications, providing essential frameworks for predictive modeling and informed decision-making. Key techniques such as regression, Bayesian inference, and likelihood estimation are pivotal within the AI domain.

Regression is a modeling technique which taken in continuous output values from input features, establishing relationships, and predicting outcomes based on historical data. Both linear and non-linear regression methods are employed depending on the complexity of the relationship between variables. In particular, non-linear regression techniques are valuable when the underlying data relationships cannot be accurately captured by a straight-line approximation. In the context of photovoltaic systems, regression models like Support Vector Machines (SVMs) and Random Forests (RF) which are inherently capable of modeling non-linear patterns have shown significant efficacy. Chaudhary et al. (2022) and Chaudhary (2019) illustrate how SVMs and RFs can predict electronic bandgaps and optimize configurations of multi-junction solar cells, effectively handling non-linear data and complex variable interactions [5; 30].

Feature Selection in Statistical Modelling plays a pivotal role in ensuring the robustness, interoperability, and interpretability of AI models. It is particularly crucial when working with high-dimensional data, where the presence of redundant, irrelevant, or noisy variables can compromise both predictive accuracy and model transparency. By selecting a subset of informative features, statistical modeling becomes more efficient and better suited for generalization beyond the training data.

A variety of feature selection techniques are employed depending on the domain and modeling objectives. Traditional methods include forward selection, which iteratively adds features that improve model performance, and backward elimination, which removes the least useful features step-by-step. More recent strategies include random permutation testing, which assesses the importance of a feature by evaluating the degradation in model performance when that feature is randomly shuffled. Additionally, Automatic Relevance Determination (ARD), often used within Bayesian

frameworks, assigns relevance weights to features during model training, automatically down-weighting those that contribute minimally to the output.

In the healthcare domain, particularly in prostate cancer diagnostics using heterogeneous MRI datasets, the Maximum Relevance Minimum Redundancy (MRMR) method has been effectively used to select radiomic features that are both informative and non-redundant. This has led to improved model performance, interpretability, and reproducibility across varied imaging conditions [16]. Complementing these selection techniques, explainability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been integrated to clarify the contribution of individual features to model predictions an essential component in clinical environments where treatment decisions are increasingly influenced by AI outputs [31].

Robust evaluation using metrics such as the Receiver Operating Curve (ROC) further supports the development of dependable AI systems. In our studies on radiomics-based machine learning models for prostate cancer detection, variability in ROC across external datasets highlighted the need for rigorous feature selection, regularization, and imputation strategies to achieve consistent performance [16]. Importantly, statistical modeling must always balance model complexity and interpretability. While complex models may offer superior performance on training data, they often do so at the cost of transparency a trade-off that is particularly consequential in high-stakes fields like healthcare.

In summary, effective feature selection whether through classical statistical techniques or modern machine learning approaches is foundational to developing AI systems that are both accurate and clinically reliable. When paired with interpretability tools, these methods ensure that models not only perform well but also align with the practical needs of end-users such as clinicians and domain experts [30; 16; 15].

2.4 Probabilistic Modelling

Probabilistic modeling plays a role in the development of adaptive and reliable AI systems, particularly in domains such as medical diagnostics and scientific modeling, where uncertainty is intrinsic. At its core, Bayesian inference provides a principled statistical framework for modeling uncertainty by updating beliefs based on new evidence. The backbone of this framework is Bayes' theorem, which enables models to integrate prior knowledge and dynamically refine predictions as new data become available [32]. This makes Bayesian methods especially suitable for environments where data are incomplete, noisy, or acquired over time.

2.4.1 Bayes' Theorem in AI Systems

Bayesian inference provides a systematic method for updating the probability of a hypothesis given new data. Mathematically, this is expressed as:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

where $P(H|D)$ is the posterior probability of hypothesis H after observing data D, $P(D|H)$ is the likelihood, $P(H)$ is the prior probability of the hypothesis, and $P(D)$ is the marginal likelihood.

In practical AI systems, this formula is operationalized in various components of the learning pipeline. For instance, in probabilistic neural networks and Bayesian Neural Networks (BNNs), the weights w of the model are treated as random variables with prior distributions. During training, Bayesian inference is used to update these weight distributions in light of observed data, leading to posterior distributions that better capture the uncertainty in model parameters. This enables the network to not only make predictions but also express confidence in those predictions a key requirement in high-stakes applications such as prostate cancer diagnosis from MRI data [33]. For example, a Bayesian convolutional neural network for lesion classification might represent its predictions as a probability distribution over diagnostic classes, thereby communicating uncertainty to the clinician and aiding decision support under ambiguity [34].

2.4.2 Bayesian Regularization in Practice

A concrete implementation of Bayesian principles in deep learning is Bayesian Regularization, which modifies the standard loss function to include a penalty term that reflects prior beliefs about model complexity. Specifically, it regularizes the sum of squared network weights:

$$L = \sum (y_i - f(x_i, w))^2 + \lambda \sum w^2$$

Here, L is the total loss, $f(x_i, w)$ is the model output, y_i is the true label, w represents the model weights, and λ controls the strength of the regularization. This approach penalizes overly complex models, thereby reducing overfitting and improving generalization. In our earlier work on optimizing silicon tandem solar cells using Artificial Neural Networks, this method enabled convergence to stable solutions under a wide range of input conditions [5].

Similarly, in medical AI models, Bayesian Regularization helps mitigate variance caused by scanner-dependent artifacts or demographic diversity. Our foundational model for prostate cancer detection employed regularization to stabilize performance across multi-vendor MRI datasets, ensuring consistent AUC performance

irrespective of imaging protocol or institution [15]. Thus, the same principle used to control overfitting in solar cell modeling proved valuable in reducing diagnostic variability in clinical imaging.

2.4.3 Comparative Advantages of Bayesian Modelling

Bayesian methods offer several advantages over frequentist and heuristic-based approaches in machine learning. Compared to traditional maximum likelihood estimation, which yields point estimates, Bayesian inference produces a full distribution over model parameters, allowing explicit uncertainty quantification. This is especially important in healthcare, where overconfident but incorrect predictions can have serious consequences. For example, a traditional deep neural network trained on a limited cohort might give high-confidence predictions for out-of-distribution cases, whereas a Bayesian model would reflect this uncertainty through widened predictive intervals.

Moreover, Bayesian models enable the seamless integration of domain expertise as prior distributions. This is less straightforward in methods like Random Forests or Support Vector Machines, which rely heavily on data-driven partitioning or kernel functions and lack a probabilistic interpretive layer. In our radiomics-based prostate cancer detection framework, the Bayesian paradigm allowed us to encode clinically relevant priors such as feature relevance derived from prior studies into the model itself, thereby enhancing both performance and transparency [16].

Bayesian approaches in probabilistic modeling serve as a robust foundation for constructing adaptive, interpretable, and generalizable AI systems. By integrating prior knowledge with observed data, these methods provide a principled way to quantify and propagate uncertainty, which is particularly essential in sensitive domains such as medical diagnostics. In our work, Bayesian inference and regularization techniques have been effectively applied to both photovoltaic optimization and clinical diagnostic tasks, supporting the development of stable, trustworthy, and high-performing AI systems across domains [5; 30; 15; 16].

2.5 Transformers

Transformer models, first introduced by Vaswani et al.[35], have emerged as a foundational architecture in artificial intelligence due to their attention-based mechanism, which enables the modeling of global dependencies in data sequences without relying on recurrence or convolution. Originally developed for natural language processing (NLP), transformers have since been extended to numerous other domains. In NLP, models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have set new benchmarks in tasks like language modeling, question answering, and text generation by effec-

tively capturing contextual information through attention-based token interactions [36; 37]. Subsequent innovations, such as XLNet and T5, have further demonstrated the scalability and adaptability of transformers in processing complex sequential data [38; 39].

Beyond NLP, the transformer architecture has been successfully applied to structured domains such as bioinformatics and molecular science, where sequence modeling is crucial [40; 41]. Most notably, the adaptation of transformers to vision tasks led to the development of the Vision Transformer (ViT), which treats images as sequences of non-overlapping patches and processes them analogously to word tokens. This approach allows the model to capture global image context more effectively than traditional convolutional neural networks (CNNs), which are typically limited to local receptive fields [42].

In this thesis, transformer-based models particularly Vision Transformers have been adapted and fine-tuned to develop computer-aided diagnostic systems for clinical imaging tasks. Unlike NLP applications, where positional embeddings capture sentence structure, Vision Transformers process spatial image information, enabling robust analysis of complex radiological patterns. For instance, in our prostate cancer detection framework, ViTs were trained on radiomic features extracted from multiparametric MRI scans, capturing inter-region relationships crucial for identifying malignancies across varied imaging protocols. These models were further optimized using domain-aware fine-tuning strategies to improve performance across multi-institutional datasets with different acquisition protocols [15].

Key components of the transformer architecture such as self-attention, positional encoding, and multi-head attention have served distinct roles in our diagnostic pipeline. The self-attention mechanism facilitated learning of long-range dependencies among radiomic features, enabling the model to consider the spatial and structural relationships between regions of interest. Positional embeddings allowed the system to preserve anatomical context, while multi-head attention enhanced the model's capacity to analyze diverse image features simultaneously. These architectural choices contributed to improved sensitivity and specificity in cancer detection, as demonstrated in our multicenter validation studies.

In summary, transformers have evolved from their initial role in NLP into a powerful framework for computer vision and medical diagnostics. In this thesis, their application to clinical imaging has enabled the development of generalizable and explainable diagnostic systems. By leveraging their architectural strengths and mitigating their limitations, transformer models have contributed to advancing the reproducibility, reliability, and clinical applicability of AI in healthcare.

2.5.1 Vision Transformers (ViT)

In medical imaging, one of the central challenges in artificial intelligence is the ability to model complex spatial dependencies and heterogeneous image features while maintaining generalizability across imaging protocols and devices. Traditional Convolutional Neural Networks (CNNs), while powerful, are constrained by their reliance on local receptive fields and limited capacity to capture global contextual relationships. These limitations often hinder diagnostic accuracy and reproducibility, especially in clinical tasks such as prostate cancer detection using MRI, where spatially distant but clinically relevant features must be interpreted together. Vision Transformers (ViTs) have emerged as a solution to this limitation by enabling global context modeling and long-range dependency learning, thereby improving diagnostic performance and generalization across diverse imaging conditions.

ViTs, initially introduced by Dosovitskiy et al., represent a paradigm shift in medical imaging by leveraging the self-attention mechanism originally developed for natural language processing [10]. Unlike CNNs that process data through convolutional kernels over local neighborhoods, ViTs treat an image as a sequence of fixed-size patches, enabling attention to be distributed dynamically across the entire image. This global attention mechanism allows ViTs to identify clinically relevant features that may be spatially distant, an essential property for capturing the full anatomical and pathological context of medical scans [10].

In medical imaging tasks, ViTs have demonstrated several advantages over CNNs, including improved scalability to higher-resolution images and better integration of global contextual information [43]. Common training strategies include pre-training on large-scale datasets followed by fine-tuning on domain-specific medical data, along with progressive image resizing and aggressive data augmentation to address limited training samples [44]. These strategies have proven particularly beneficial in radiomics applications, such as prostate cancer detection from MRI. ViTs enable automatic extraction of high-level textural and morphological features that are difficult for human experts to annotate and for CNNs to capture explicitly [45].

In our work, ViTs have been applied to learn discriminative representations from multiparametric MRI scans for the purpose of identifying prostate cancer. These models demonstrated improved sensitivity and specificity by effectively distinguishing between cancerous and non-cancerous tissues, benefiting from the transformer's capacity to learn long-range spatial dependencies and integrate global image features [15]. For instance, our foundational radiomics model using ViTs achieved an AUC of 0.86 on heterogeneous datasets, which improved to over 0.90 following fine-tuning with site-specific data [15].

A key strength of ViTs in clinical imaging lies in their cross-site generalizability. Variations in MRI acquisition protocols across institutions often lead to distributional shifts, reducing the reliability of standard models. Fine-tuning ViTs on small, site-

specific datasets after pre-training on large, heterogeneous cohorts has proven effective in mitigating this issue [46]. Moreover, harmonization strategies and the use of foundational models have been instrumental in ensuring consistent performance and reproducibility in clinical environments.

The adoption of explainable AI techniques, such as SHAP and LIME, further enhances the clinical trustworthiness of ViT-based systems. By identifying which patches or features most influence the model's decisions, these tools help clinicians interpret predictions and validate their alignment with medical understanding [15]. This transparency is crucial for integration into routine clinical workflows, where decision support tools must be both accurate and interpretable.

In summary, Vision Transformers address fundamental challenges in medical imaging AI by offering improved context modeling, cross-site generalization, and interpretability. Their successful application in prostate cancer detection illustrates their potential as core components of clinical decision support systems. Future research will likely explore multimodal transformer architectures, enhanced interpretability mechanisms, and integration with real-time clinical diagnostics to further advance their utility in healthcare.

2.5.2 Attention Maps

Attention mechanisms have revolutionized the field of deep learning by enabling models to dynamically focus on specific parts of the input data that are crucial for a given task. This capability enhances the interpretability and performance of neural networks across various domains, including natural language processing (NLP), computer vision, and speech recognition. Attention mechanisms provide a way for models to allocate computing resources towards the most informative components of the input data, thus improving efficiency and effectiveness [47]. Initial neural network architectures, such as feed-forward and recurrent neural networks, processed inputs using a static architecture without the ability to focus selectively on parts of the input. The introduction of attention mechanisms allowed these models to mimic a more human-like focus and processing, adapting dynamically to the input based on the context provided by other parts of the data.

Developed by Bahdanau et al., this attention mechanism introduces a way to align and translate different parts of an input sequence into an output sequence [48]. The core idea relies on calculating a context vector as the weighted sum of hidden states, where the weights are derived by a trainable alignment model. This model scores how well the inputs around position j and the output at position t match:

$$\text{Context}_t = \sum_j \alpha_{tj} h_j, \tag{2.3}$$

where,

$$\alpha_{tj} = \frac{\exp(\text{score}(s_{t-1}, h_j))}{\sum_k \exp(\text{score}(s_{t-1}, h_k))} \tag{2.4}$$

The Transformer model, introduced by Vaswani et al., utilizes self-attention where the alignment score is calculated between different positions of a single sequence, to represent the sequence through its internal dependencies [35]. The key equations involve queries, keys, and values derived from the input vector, using different learned linear transformations:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.5}$$

where Q, K, V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the keys.

Attention maps are visual representations that illustrate the areas where the model is focusing its attention within the input data. These maps are crucial for interpreting the decision-making process of the model, making the model's predictions more transparent and trustworthy.

Attention maps can be generated by plotting the attention weights α_{tj} from models like Bahdanau attention or the softmax outputs from Transformer models [48]. These maps are particularly useful in tasks like image recognition, where they highlight the image regions influencing the model's predictions. In NLP, attention mechanisms help models focus on relevant words or phrases in tasks like machine translation and sentiment analysis. For example, when translating a sentence, the model can focus more on subject-object relationships, improving the quality of the translation [49].

In computer vision, attention maps are used to identify regions of interest in images, significantly improving tasks such as object detection and classification. For instance, attention-guided models have enhanced the detection of defects in solar panels by focusing on anomalous regions [50]. In speech recognition, attention mechanisms enable models to focus on specific parts of an audio signal, improving transcription accuracy by aligning audio features with textual outputs effectively. Attention mechanisms significantly enhance the performance, robustness, and explainability of AI models. By allowing models to focus selectively on the most informative parts of the data, they not only improve the accuracy but also bring transparency to the decision-making process, which is crucial for applications in fields requiring reliable and interpretable outcomes. The continued development and application of attention mechanisms across different domains promise further advancements in the capabilities of AI systems.

2.6 Summary of the Identified Research Gap

From the comprehensive review of statistical, probabilistic, and transformer-based methods in artificial intelligence presented in this chapter, it is evident that despite significant advancements, critical challenges persist concerning model generalizability, reproducibility, and interpretability, particularly in heterogeneous, real-world clinical and photovoltaic datasets. While current statistical and probabilistic approaches effectively quantify uncertainty and incorporate domain knowledge, their robustness across diverse imaging protocols, scanner vendors, and environmental conditions remains limited. Transformer-based models, such as Vision Transformers (ViTs), have shown promise in capturing global contextual dependencies, however, their performance heavily depends on careful fine-tuning and harmonization strategies, especially when applied to small, diverse, or multi-institutional datasets. Moreover, existing methods lack comprehensive integration of multimodal data streams, rigorous cross-vendor validation protocols, and explicit uncertainty quantification mechanisms, all crucial for ensuring safe and reliable deployment in high-stakes domains like medical diagnostics and energy optimization. This thesis specifically addresses these identified gaps by developing foundational AI models enhanced with probabilistic reasoning, explainable AI techniques, and cross-domain generalization strategies, aiming to improve reproducibility, interpretability, and performance across diverse clinical and photovoltaic applications.

3 Datasets and Data Preprocessing

3.1 Importance of Data Quality and Preprocessing

The successful application of artificial intelligence (AI) in domains such as radiomics and photovoltaics is fundamentally dependent on the quality of data and the rigor of preprocessing strategies applied to it. Data preprocessing is not solely aimed at improving adaptability, rather, it encompasses a broad set of techniques including noise reduction, normalization, scaling, encoding, and outlier handling. These steps are essential to ensure data consistency, enhance model learning, and improve both the interpretability and performance of AI systems across diverse datasets [5; 15].

In both medical imaging and photovoltaic research, datasets often exhibit significant variability due to differences in imaging protocols or experimental conditions. For example, in radiomics, variability across MRI vendors introduces heterogeneity that can compromise model generalizability [46]. Similarly, in photovoltaics, data derived from solar cell simulations must be carefully curated to ensure consistency in spectral power distribution, material layer properties, and temperature conditions. In our solar cell optimization studies, simulation data were generated through over one million iterations using SCAPS-1D software, where inputs such as spectral density, temperature, and layer thicknesses were systematically varied. Ensuring the correctness and representativeness of these simulation conditions was critical for training reliable neural networks [5].

3.1.1 Data Quality and Domain-Specific Considerations

Data quality directly affects the robustness and generalizability of AI models. High-quality datasets are typically characterized by completeness, accuracy, consistency, and relevance. In radiomics, for instance, low-quality image inputs or incomplete segmentations can lead to extraction of unreliable features and degraded diagnostic performance. Similarly, in photovoltaic simulation data, ensuring precise parameter definitions (e.g., uniform layer thicknesses and accurate temperature profiles) is essential for trustworthy model predictions.

In both domains, quality control is enforced through domain-specific validation metrics. In radiomics, signal-to-noise ratio, inter-scanner reproducibility, and segmentation fidelity are assessed. In solar cell modeling, consistency in I-V charac-

teristics, convergence of simulated results, and accurate reproduction of material behavior under different spectral inputs are considered indicators of data integrity. These considerations ensure that models trained on such data reflect real-world performance and are not biased by artifacts or inconsistencies.

3.1.2 Preprocessing Pipelines

An effective preprocessing pipeline is necessary to align diverse data sources with the requirements of machine learning algorithms. Preprocessing ensures that the data fed into models is consistent, reliable, and representative of the phenomena being studied. In radiomics, this pipeline typically includes steps such as image resampling, normalization, intensity standardization, and segmentation. Additional essential components include denoising, to suppress scanner-induced artifacts, and outlier removal, to eliminate anomalous feature values that may arise from segmentation errors or imaging inconsistencies. These steps are crucial for maintaining anatomical fidelity and ensuring that downstream features are both reproducible and clinically interpretable.

In contrast, preprocessing for solar cell optimization involves organizing high-dimensional spectral data, encoding structural parameters (such as layer thicknesses and temperature gradients), and normalizing physical measurements (e.g., current density, voltage) before feeding them into neural networks [5; 51]. Here too, outlier detection is important, especially when dealing with simulated datasets generated across wide parametric sweeps, where numerical instabilities or edge cases can introduce noise or invalid data points. Although methods such as Bayesian Regularization and feature transformation have been discussed earlier in the context of model optimization, their practical integration into preprocessing workflows goes beyond theoretical formulation. Domain expertise is indispensable in identifying which attributes of raw data require transformation or correction. For example, in radiomics, careful handling of spatial orientation during normalization ensures that anatomical context is preserved across patients. In photovoltaics, preprocessing must ensure that feature transformations do not distort the physical relationships among simulation variables, which could undermine the fidelity of performance predictions.

3.2 Variable Selection and Feature Engineering

In high-dimensional domains such as medical imaging and solar cell simulation, the selection of relevant variables and the design of informative features are key to developing models that are both accurate and generalizable. Feature selection serves to eliminate redundant or irrelevant information, reduce overfitting, and improve computational efficiency. To maintain structural coherence, earlier references to feature selection such as MRMR have been limited to brief mentions, with detailed discus-

sion provided here.

3.2.1 Prominent Feature Selection Methods

The details of all the state-of-the-art feature selection methods have also been consolidated in Table:3.1 with strength and limitations. Listing them as below:

- Maximum Relevance Minimum Redundancy (MRMR): Selects features that are maximally relevant to the target variable and minimally redundant with each other.
- Recursive Feature Elimination (RFE): Recursively removes less important features based on model performance until an optimal subset is identified.
- LASSO (Least Absolute Shrinkage and Selection Operator): Performs feature selection through regularization by shrinking some coefficients to zero.
- ReliefF: Evaluates the relevance of features based on how well their values distinguish between near instances.

Each of these methods has trade-offs. MRMR is well-suited for noisy datasets and was used in our prostate cancer models to select stable radiomic features across imaging vendors [46]. In contrast, LASSO provides strong regularization benefits but can struggle with correlated features. RFE is often used with tree-based models but is computationally intensive, especially on large feature sets.

3.2.2 Domain-Specific Feature Engineering

Feature engineering transforms raw data into representations that improve model learning. In radiomics, this includes intensity-based features (mean, variance), texture features (entropy, contrast), and shape-based metrics, often extracted using tools like PyRadiomics or MRCradiomics [52]. Our pipeline utilized these tools to generate a high-dimensional feature matrix from multiparametric MRI scans, which were then filtered using MRMR to train discriminative models [46].

In photovoltaics, feature engineering centers around parameters such as spectral power density, temperature, and semiconductor layer thicknesses. These parameters were varied systematically in simulations and then normalized to ensure numerical stability in neural network training. The engineered feature matrix enabled the neural network to predict key outcomes such as open-circuit voltage, fill factor, and maximum power point [30; 5].

Together, feature selection and engineering form the foundation of model development. Their correct implementation ensures that the trained models are not only accurate but also robust to data heterogeneity and adaptable across clinical or experimental conditions.

Table 3.1. Comparison of Prominent Feature Selection Methods

Method	Selection Principle	Strengths	Limitations
MRMR	Maximizes relevance to the target variable while minimizing redundancy among features	Well-suited for small and noisy datasets, emphasizes informative and non-redundant features	May overlook interactions between features
RFE	Recursively removes the least important features based on model performance	Provides a model-specific optimal subset of features	Computationally intensive, especially on high-dimensional data
LASSO	Applies L1 regularization to shrink some feature coefficients to zero	Performs both feature selection and regularization, useful for interpretability	Can struggle with correlated variables, choice of regularization strength is critical
ReliefF	Assigns weights to features based on instance-based relevance using similarity and label information	Captures local feature dependencies and interactions	Sensitive to noisy or imbalanced datasets, dependent on sample size

3.3 Radiomics Dataset for Prostate Cancer Detection

3.3.1 Multicenter Data Collection and Ethics

The multicenter MRI data (MULTI-IMPROD) collection process for prostate cancer detection involved an extensive collaboration between renowned institutions in Finland, and USA [53]. This collaborative effort aimed to capture a diverse dataset critical for enhancing the robustness and generalizability of AI models in prostate cancer detection. Ethical approvals were meticulously secured from the various Institutional Review Boards (IRBs) of each participating institution, ensuring that the study adhered strictly to ethical standards and maintained patient confidentiality. The informed consent process was implemented with the utmost care, providing participants comprehensive information regarding the study's objectives, potential risks, and benefits. This ethical rigor is crucial, as it underscores the importance of patient autonomy and respect in clinical research. The strategic inclusion of cross-vendor and cross-institutional data diversity in the dataset was paramount. It allowed for the creation of a foundation upon which robust AI models could be built, ensuring these models are not only accurate but also adaptable to different clinical settings and imaging conditions [15].

3.3.2 Image Pre-Processing and Radiomic Feature Extraction

In the domain of prostate cancer detection using MRI data, image preprocessing is a critical step that directly impacts the accuracy, reproducibility, and clinical validity of radiomic analysis. The preprocessing pipeline is designed to enhance the quality and consistency of image data extracted from T2-weighted and diffusion-weighted MRI scans, which are commonly used in prostate imaging due to their high sensitivity to anatomical and cellular-level changes.

A conventional clinical approach often involves calculating mean Apparent Diffusion Coefficient (ADC) values within the manually segmented Region of Interest (ROI). While this method provides a quick quantitative summary of diffusion characteristics, it may overlook important heterogeneity within the lesion and surrounding tissue. Thus, although mean ADC values remain a clinically accepted and widely used metric, they serve as a reference baseline when evaluating the added value of radiomic features in advanced AI-based analysis.

Region of Interest (ROI) Segmentation

ROI segmentation is the foundational step in preparing MRI data for radiomic analysis. The objective is to delineate the prostate and any potential lesions, ensuring that the features extracted are relevant to the disease being studied. Manual delineation, automated methods, or semi-automated algorithms can be used, each with its own

advantages and limitations [54]. Recent advancements employ deep learning models for improved accuracy and consistency in segmentation [55]. Mathematically, segmentation can be represented as a function $S : \mathbb{R}^n \rightarrow \{0, 1\}$ that assigns each voxel in the image I to either the ROI or background, where n represents the image dimensions.

Intensity Normalization

To account for variability in image intensity, which may arise from differences in MRI scanner settings, acquisition protocols, or patient-specific physiological factors, intensity normalization is a crucial preprocessing step. This variability can negatively affect the performance of AI models by introducing inconsistencies in voxel intensity distributions, leading to biased feature extraction and poor generalization across datasets. Normalization ensures that the intensity scale is harmonized across all input scans, thereby enhancing the reliability and reproducibility of radiomic features and improving downstream machine learning performance [56].

Common techniques for intensity normalization include Z-score normalization and min-max scaling. Z-score normalization, in particular, transforms intensity values to a standardized scale with zero mean and unit variance:

$$I_{norm} = \frac{I - \mu}{\sigma}$$

where I represents the raw intensity, μ is the mean intensity, and σ is the standard deviation within the region of interest. This approach ensures comparability of voxel intensities across subjects and scanning sessions, facilitating consistent radiomic feature extraction and aiding the robustness of predictive models.

Transformation Methods

To ensure anatomical consistency and alignment across images from different sessions or sequences, geometric transformations are applied during preprocessing. These transformations correct for variations in patient positioning and organ orientation, which is essential for extracting spatially consistent radiomic features from corresponding regions across scans.

The most commonly applied transformations are affine transformations, which include translation, rotation, scaling, and shearing. These are typically represented using 4×4 transformation matrices in 3D space:

$$T(\mathbf{x}) = R \cdot \mathbf{x} + \mathbf{t}$$

where R is the rotation matrix, $\mathbf{t}$ is the translation vector, and additional scaling factors can be incorporated as needed. Affine transformations maintain paral-

lelism and straight lines, making them suitable for aligning global structures like the prostate gland.

However, in cases where local anatomical deformation needs to be accounted for such as soft tissue displacement between sessions or patient-specific variations non-rigid transformations are employed. These are often modeled using B-spline functions, thin-plate splines, or optical flow-based deformation fields. Non-rigid registration enables finer alignment of tissue structures, particularly when analyzing subtle morphological or textural features that are critical for prostate cancer detection.

By integrating both affine and non-rigid transformations, the preprocessing pipeline ensures that the extracted features are spatially consistent, anatomically meaningful, and suitable for robust model training across heterogeneous imaging datasets.

Radiomic Feature Extraction

Once preprocessing is complete, radiomic features are extracted using standardized protocols. This involves computing various textures, shape, and intensity-based features from the segmented ROIs. The radiomics analysis typically utilizes libraries such as Pyradiomics, which implements standardized feature definitions, ensuring reproducibility across studies [57]. Formally, the feature extraction process can be described using mathematical descriptors such as the Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Run-Length Matrix (GLRLM), and others. The GLCM, for example, is defined as a matrix $P(i, j)$, where each element represents the frequency of co-occurrence of pixel pairs with intensities i and j at a specified distance and orientation [58]. The features derived from GLCM include contrast, correlation, energy, and homogeneity, each providing a distinct texture characteristic.

3.3.3 Pyradiomics and MRCradiomics

Pyradiomics is a widely recognized open-source platform, lauded for its extensive feature set and adherence to Image Biomarker Standardization Initiative (IBSI) guidelines, which emphasizes the reproducibility and robustness of computed features [59]. It supports the extraction of a comprehensive range of features, including first-order statistics, shape descriptors, textural features (e.g., GLCM, GLRLM), and higher-order features derived from filter transformations. MRCradiomics, in contrast, extends the capabilities of Pyradiomics by incorporating advanced algorithms tailored for MR images. It includes proprietary methods for optimizing feature extraction, particularly under variability introduced by different MRI vendors [60]. MRCradiomics excels in cross-vendor reproducibility, a crucial factor given the variability reported in MRI-based studies as highlighted by recent research [61].

Categories of Radiomic Features Both packages compute a range of radiomic features crucial for cancer classification:

- **GLCM Features**: These include contrast, dissimilarity, homogeneity, ASM (angular second moment), energy, and correlation. GLCM features are vital in capturing textural properties that distinguish between aggressive and non-aggressive cancer tissues.
- **GLRLM Features**: These features measure the length of consecutive runs of pixels with the same gray level, such as short-run emphasis, long-run emphasis, and gray-level non-uniformity, helping to characterize the heterogeneity of tissue.
- **Shape Features**: These are geometric descriptors capturing the 3D morphology of the tumor, including volume, surface area, compactness, and sphericity. Accurately quantifying tumor shape aids in the assessment of tumor aggressiveness and potential invasiveness.

The reproducibility and robustness of radiomic features are paramount for clinical application. Both Pyradiomics and MRCradiomics emphasize standardized feature extraction protocols to ensure that analyses are not only repeatable within the same dataset but also generalizable across different scanning environments. This robustness is critical given the stark differences in feature performance across vendors, as demonstrated in studies where combined features from Pyradiomics and MRCradiomics were shown to enhance diagnostic accuracy [62].

In conclusion, the selection and implementation of the appropriate radiomic toolkit can significantly influence the outcome in prostate cancer detection. Both Pyradiomics and MRCradiomics offer valuable insights, but the choice between them should be guided by specific requirements regarding feature robustness, dataset compatibility, and vendor variability, tailored to clinical diagnostic needs.

3.4 Data Preprocessing for Radiomics Research

In the context of machine learning, feature selection is a crucial preprocessing step that enhances model performance and interpretability by identifying the most relevant features in relation to the target variable. Among the widely used metrics for this purpose are Information Gain (IG) and Mutual Information (MI), both of which assess the dependence between input features and the target variable based on information theory. Information Gain quantifies the reduction in entropy (uncertainty) about the target variable when the value of a feature is known, while Mutual Information measures the amount of information shared between a feature and the target, treating their relationship symmetrically.

Mathematically, Information Gain for a feature X and target variable Y is expressed as:

$$IG(Y, X) = H(Y) - H(Y|X)$$

where $H(Y)$ is the entropy of the target variable and $H(Y|X)$ is the conditional entropy of Y given X. The entropy $H(Y)$ and conditional entropy $H(Y|X)$ are computed as:

$$H(Y) = -\sum_{y \in Y} P(y) \log_2 P(y), \quad H(Y|X) = -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \log_2 P(y|x)$$

[63].

Mutual Information is given by:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right)$$

[64].

These metrics are particularly useful in classification problems, where the objective is to determine which features are most informative for predicting a class label. For instance, in medical imaging applications such as prostate cancer detection, radiomic features derived from MRI scans can be ranked using IG or MI with respect to the disease outcome, helping to identify those features most relevant for training predictive models [65].

However, despite their theoretical appeal, both IG and MI come with practical challenges. A key issue lies in the estimation of joint and conditional probability distributions, particularly in high-dimensional feature spaces where the number of possible value combinations grows exponentially. Discretization of continuous variables often required for these metrics can introduce bias and lead to loss of information. Moreover, both IG and MI evaluate features independently and thus do not account for feature interactions or redundancy among correlated features. This can result in the selection of multiple features that individually appear important but contribute overlapping information, which may not improve and could even degrade model performance.

To address these limitations, IG and MI are often combined with other methods, such as redundancy analysis (e.g., in Maximum Relevance Minimum Redundancy, MRMR), forward selection strategies, or embedded techniques like Automatic Relevance Determination (ARD), which account for inter-feature dependencies during the selection process.

In practice, after calculating IG or MI scores, features are ranked, and those with higher values are retained for model training. This helps to reduce dimensionality,

improve generalization by minimizing overfitting, and enhance computational efficiency during model development [66].

3.4.1 Pipeline of MRMR and Information Gain

The integration of Minimum Redundancy Maximum Relevance (MRMR) with Mutual Information creates a robust pipeline for feature selection, enhancing the stability and reproducibility of machine-learning models. MRMR aims to select features that have maximum relevance to the target variable and minimum redundancy among themselves, which is particularly important in domains with highly correlated features, such as radiomics. The pipeline begins with preprocessing steps, including data normalization and handling missing values to ensure consistency across features. Initial feature selection is performed using Mutual Information to assess the relevance of each feature to the target variable. Features surpassing a predefined threshold of MI are retained for further analysis [67]. Subsequently, MRMR is applied to filter out redundant features among those selected in the first step. The relevance of a feature is evaluated against its redundancy with previously selected features using the criterion:

$$\max_{S} \left[\frac{1}{|S|} \sum_{x_i \in S} MI(x_i, Y) - \frac{1}{|S|^2} \sum_{x_i, x_j \in S} MI(x_i, x_j) \right] \tag{3.1}$$

where S is the subset of selected features, $MI(x_i, Y)$ is the mutual information between feature x_i and the target Y, and $MI(x_i, x_j)$ is the mutual information between pairs of features (x_i, x_j).

The threshold values for MI and the criteria for MRMR are set based on domain-specific knowledge and empirical analysis to ensure an optimal balance between relevance and redundancy [68]. Parameters are chosen based on cross-validation results to maximize the model's Area Under the Curve (AUC) and to ensure stability across different datasets. This pipeline's use stabilizes the model training phase by reducing the feature space's dimensionality without sacrificing informative content. By focusing on relevant and non-redundant features, models can achieve greater robustness across multiple sites, as demonstrated by improved cross-vendor reproducibility in computer-aided diagnosis models for prostate cancer detection [69]. The pipeline's ability to maintain model performance consistency is crucial in clinical applications where reproducibility is paramount for reliable decision support.

In summary, integrating MRMR with Mutual Information in a structured pipeline offers substantial benefits for feature selection. This method not only enhances the generalizability of AI models across diverse datasets but also supports their robustness and reliability in practical applications.

3.5 Photovoltaic Dataset

3.5.1 Dataset Description and Characteristics

This section provides a comprehensive description of the photovoltaic datasets employed in this dissertation, specifically focusing on their origins, formats, and critical attributes. The datasets are pivotal in the analysis and optimization of photovoltaic cell performance, with key variables including spectral power density, temperature, and material properties such as thickness and band gap. Furthermore, this dissertation extends to examining the challenges posed by experimental variability and representational sparsity, underscoring the significance of statistical rigor and mathematical analysis to mitigate these issues.

The photovoltaic datasets utilized in this dissertation originate from a variety of sources, encompassing experimental data from laboratory settings and simulation outputs from sophisticated modeling software such as SCAPS-1D. The data format is typically structured in tabular form, where each entry corresponds to a unique configuration or measurement condition. Parameters recorded include spectral power density, ambient and operational cell temperatures, and physical properties of the materials, such as layer thickness and band gap energy. For instance, the research conducted by Chaudhary et al. (2019) on multi-junction solar cells incorporated data generated from SCAPS-1D simulations to analyze their quantum efficiency and fill factor characteristics. Additionally, the study involving silicon tandem solar cells leveraged artificial neural networks (ANNs) to optimize input parameters such as spectral data, power density, and temperature, aligning physical measurements with computational predictions.

The analysis of photovoltaic datasets necessitates a detailed understanding of several critical variables:

- Spectral Power Density: This variable represents the distribution of power across different wavelengths of light incident on the solar cell. It is a crucial factor in determining the efficiency of light absorption and subsequent energy conversion.
- Temperature: Both ambient and operational temperatures are measured, as these significantly impact the performance and efficiency of photovoltaic cells. Higher temperatures generally lead to increased carrier recombination rates, affecting overall efficiency.
- Material Properties: Essential material properties include the thickness of each solar cell layer and the band gap energy, which defines the energy threshold for electron excitation, pivotal in determining the photovoltaic efficiency.

These variables collectively determine the input space for the simulation or modeling process and are critical in the accurate prediction and optimization of photovoltaic

performance across diverse material and environmental configurations.

4 Optimization of Neural Networks

Artificial Intelligence (AI) has achieved remarkable empirical success across numerous domains, ranging from image classification to medical diagnosis. However, the theoretical understanding of optimization dynamics, particularly in deep learning, remains incomplete. One of the primary challenges in neural network training is the non-convexity of the loss function landscape, which gives rise to saddle points, local minima, and unstable trajectories.

To address these challenges, this chapter introduces a principled framework that integrates dynamic learning rates, superlevel set analysis, and Lyapunov stability theory. The central objective is to construct a mathematically grounded approach that ensures stable convergence and generalization by analyzing the topology of loss functions and the dynamics of training algorithms.

This framework not only facilitates a deeper understanding of optimization dynamics but also informs the design of adaptive learning schedules and robust cost functions. A key novelty lies in demonstrating the connectedness and stability of superlevel sets under exponential learning rate decay, which provides both theoretical guarantees and practical performance benefits.

4.1 Challenges in Neural Network Optimization

Neural networks are trained by minimizing a loss function $L(\theta)$, where $\theta \in \mathbb{R}^n$ denotes the parameter vector. The high-dimensional and non-convex nature of L makes the optimization landscape intricate. Specifically, the following challenges are commonly encountered:

- **Saddle Points:** Stationary points where the gradient vanishes but are neither minima nor maxima, impeding gradient-based methods.
- **Poor Local Minima:** While rare in high dimensions, they can still affect convergence.
- **Exploding or Vanishing Gradients:** Particularly prominent in deep networks, these issues destabilize the optimization process.
- **Overfitting:** Occurs when the model fits training data excessively, degrading generalization.

Traditional optimization algorithms, such as Stochastic Gradient Descent (SGD) with static learning rates, often struggle to adapt to these dynamics. Recent methods utilize learning rate schedules or adaptive optimizers, but a rigorous theoretical basis for their effectiveness remains underdeveloped.

4.2 Superlevel Sets in Neural Optimization

Let us define a superlevel set of a loss function $L : \mathbb{R}^n \to \mathbb{R}$ as:

$$S_\lambda = \{\theta \in \mathbb{R}^n : L(\theta) \geq \lambda\}, \quad \lambda \in \mathbb{R} \tag{4.1}$$

Superlevel sets encapsulate the regions in parameter space where the loss exceeds a specified threshold λ. The structure of these sets reflects important geometric properties of the loss landscape, such as the existence of barriers between basins of attraction and the stability of descent paths.

A crucial property in this context is connectedness. If the superlevel set S_λ is connected, then there exists a continuous path within S_λ linking any two points $\theta_1, \theta_2 \in S_\lambda$. This ensures that optimization trajectories do not fall into disconnected subregions, a desirable property for ensuring stable convergence.

in the next section, we will show that under dynamic learning rates and certain Lyapunov conditions, superlevel sets remain connected across training iterations.

4.3 Lyapunov Stability Framework

We model neural network training as a discrete-time dynamical system with state vector θ_t. A loss function $L(\theta)$ acts as a Lyapunov candidate function. For a continuous-time system, the Lyapunov stability condition is:

$$\frac{dV}{dt} \leq 0 \quad \text{for all } t \geq 0 \tag{4.2}$$

In our discrete setting, using gradient descent with learning rate $\eta(t)$, the update rule is:

$$\theta_{t+1} = \theta_t - \eta(t)\nabla L(\theta_t) \tag{4.3}$$

We choose $V(\theta) = L(\theta)$. Then,

$$V(\theta_{t+1}) - V(\theta_t) \approx -\eta(t)\|\nabla L(\theta_t)\|^2 \tag{4.4}$$

We choose $V(\theta) = L(\theta)$, where V represents the total energy quantity of the system, and its decrease indicates progress towards equilibrium or convergence, hence denoting a healthy convergence. Then,

$$V(\theta_{t+1}) - V(\theta_t) \approx -\eta(t)\|\nabla L(\theta_t)\|^2 \tag{4.5}$$

If $\eta(t) > 0$, the Lyapunov function is guaranteed to decrease monotonically, ensuring stability. Furthermore, under an exponentially decaying learning rate:

$$\eta(t) = \eta_0 e^{-\alpha t}, \quad \alpha > 0 \tag{4.6}$$

the decay introduces a natural transition from exploration to exploitation, allowing wide parameter searches early in training and finer updates in later stages.

4.4 Connectivity of Superlevel Sets Under Lyapunov Stability

We aim to show that under the assumption:

$$\nabla V(\theta) \cdot \nabla L(\theta) \geq 0 \quad \forall \theta \in \mathbb{R}^n \tag{4.7}$$

the superlevel sets S_λ are connected for all λ.

Theorem: Let $L(\theta)$ be continuously differentiable and $V(\theta) = L(\theta)$ be a Lyapunov function satisfying the above condition. Then for any λ, the superlevel set S_λ is connected.

Proof Sketch: Since the gradient descent update maintains $L(\theta_{t+1}) \leq L(\theta_t)$, the optimization path remains within S_λ if $\theta_0 \in S_\lambda$. The inner product condition guarantees that the flow induced by the gradient descent direction never moves away from S_λ, preserving connectedness across iterations.

This property ensures that optimization does not fragment into isolated basins, which enhances robustness to initialization and improves convergence guarantees.

4.5 Bayesian Perspective on Learning Dynamics

The training process can be framed as maximizing the posterior:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{4.8}$$

Taking the negative log posterior:

$$\mathcal{L}(\theta) = -\log P(D|\theta) - \log P(\theta) \tag{4.9}$$

The gradient descent update rule becomes:

$$\theta_{t+1} = \theta_t - \eta(t)\nabla_\theta \mathcal{L}(\theta_t) \tag{4.10}$$

The exponential decay of $\eta(t)$ ensures that early iterations prioritize likelihood (data fitting), while later stages incorporate stronger influence from the prior, improving generalization and preventing overfitting.

4.6 Dynamic Cost Function and Regularization

To formalize the robustness and adaptability of training, we define the dynamic cost function:

$$J_{\text{dynamic}}(\theta; D, t) = \gamma(t) \left[\frac{1}{N} \sum_{i=1}^{N} w_{y_i} \rho(y_i, x_i) L(y_i, f(x_i; \theta)) + \lambda \Omega(\theta) \right] \tag{4.11}$$

Where:

- $\gamma(t) = 1 + \kappa e^{-\delta t}$ is a temporal modulation factor,
- w_{y_i} addresses class imbalance,
- $\rho(y_i, x_i)$ discounts noisy samples,
- $\Omega(\theta)$ is a regularization term (e.g., ℓ_2 norm).

The combination ensures early flexibility and late-stage regularization, maintaining stability across complex data distributions.

4.7 Gradient Descent with Exponentially Decaying Learning Rate

Let us analyze the behavior of the loss after one gradient update:

$$L(\theta_{t+1}) \approx L(\theta_t) - \eta(t) \|\nabla L(\theta_t)\|^2 \tag{4.12}$$

Using $\eta(t) = \eta_0/(1 + \|\nabla L(\theta_t)\|)$, we obtain:

$$L(\theta_{t+1}) \leq L(\theta_t) - \frac{\|\nabla L(\theta_t)\|^2}{1 + \|\nabla L(\theta_t)\|} \tag{4.13}$$

This ensures that the loss consistently decreases, confirming convergence. Moreover, it avoids large jumps near steep gradients and ensures fine adjustments in flatter regions.

The proposed framework provides a theoretical foundation for designing learning rate schedules that guarantee stable convergence. It also offers insights into robust cost function design and principled regularization.

4.8 Bringing the Framework Together

This chapter introduced several core concepts - dynamic learning rates, superlevel sets, Lyapunov stability, Bayesian learning dynamics, and a time-dependent cost function that each play a specific role in improving neural network optimization.

However, these ideas are not intended as disconnected observations. Instead, they form an integrated framework that builds toward one goal: developing training dynamics that are not only stable and convergent but also adaptable and robust across complex, high-dimensional tasks.

While each individual concept (e.g., exponential decay, Lyapunov stability, Bayesian interpretation) has been studied before, the novelty of this work lies in how these components are mathematically unified. The chapter shows, for the first time, how exponential decay directly supports connected superlevel sets under Lyapunov assumptions an idea not commonly formalized. Furthermore, the notion of equiconnectedness introduced here provides a new way to understand optimization stability across iterations.

Moreover, the design of the dynamic cost function which includes a time-weighting term $\gamma(t)$, label-specific weights w_y, and a noise discount factor $\rho(x, y)$ is novel in how it explicitly evolves with training time, adapting to the current state of learning.

An example of how it would work in practice is given below:

- Robustness to noisy labels: If certain samples are mislabeled or difficult to classify (e.g., ambiguous MRI slices or edge-case photovoltaic simulations), the $\rho(x_i, y_i)$ term in the dynamic cost function reduces their influence during training. This prevents the model from overfitting to unreliable samples.

- Adaptability to class imbalance: In many real-world datasets, certain classes appear more frequently than others. The weight term w_{y_i} gives more importance to underrepresented classes, allowing the model to learn a balanced decision boundary.

- Gradual regularization: The time-decaying factor $\gamma(t)$ starts high, allowing the model to learn freely in the early stages. As training progresses, $\gamma(t)$ shrinks, giving more importance to the regularization term $\Omega(\theta)$. This helps control model complexity and supports better generalization.

5 Clinical Decision Support System (CDSS)

5.1 Introduction

In recent years, foundational models have emerged as a transformative paradigm in artificial intelligence (AI), particularly for tasks involving high-dimensional data and complex feature interactions. These models, often characterized by their ability to scale across diverse tasks through transfer learning and fine-tuning, exhibit robust generalization capabilities when trained on extensive and heterogeneous datasets. Foundational models serve not only as initial baselines for downstream tasks but also as adaptable engines that can be repurposed with minimal data and effort for domain-specific applications. Their architecture frequently built upon transformers or other deep neural structures enables contextual feature aggregation, thereby preserving both global and local representations critical to downstream performance. Within the context of medical imaging, the deployment of foundational models facilitates the standardization of diagnostic tools by reducing site-specific bias and enhancing reproducibility. This is particularly relevant in radiomics, where variations in imaging protocol, scanner vendor, and patient demographics often confound model performance. By pretraining on heterogeneous datasets and adopting a modular design, foundational models accommodate diverse input distributions and support fine-tuning strategies tailored to site-specific constraints. This aligns with current clinical needs, where diagnostic systems must maintain reliability across centers while adapting to local contexts without exhaustive retraining.

Radiomics features, often imperceptible to the human eye, characterize the shape, texture, intensity, and spatial relationships within defined regions of interest (ROIs). The process begins with standardized image acquisition and segmentation, followed by the extraction of features using mathematical transformations and statistical computations. Features derived from platforms such as PyRadiomics and MRCRadiomics include, but are not limited to, first-order statistics (e.g., mean, variance, skewness), shape-based descriptors (e.g., sphericity, volume), and texture-based matrices such as the Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), and Gray Level Run Length Matrix (GLRLM). Transform-based descriptors such as wavelet decompositions and Laplacian of Gaussian filtered textures capture multiscale representations of tumor heterogeneity. Moreover, custom-

built feature extraction scripts in MRCRadiomics offer domain-specific texture measures such as Zernike moments and Frangi filters, which enrich the feature pool for specialized applications like prostate cancer detection. The calculated features serve as inputs to machine learning pipelines where they are filtered using statistical and information-theoretic techniques such as Mutual Information (MI), Information Gain (IG), and Maximum Relevance Minimum Redundancy (MRMR). These selection methods ensure the retention of informative, non-redundant predictors while mitigating overfitting. Radiomics thus transforms qualitative image interpretation into a quantitative, reproducible, and scalable workflow capable of supporting high-performance clinical decision support models.

Clinical Decision Support Systems (CDSS) are algorithmically driven platforms that assist healthcare professionals in making evidence-based diagnostic and therapeutic decisions. In oncology, these systems integrate multimodal data ranging from genomics and pathology to radiological images to inform diagnosis, risk stratification, treatment planning, and monitoring. Radiomics-based CDSS models focus on the analysis of medical imaging data, where the integration of radiomic features with AI models has shown promise in augmenting radiologist performance and reducing diagnostic variability.

In the domain of prostate cancer, CDSS models trained on T2-weighted and diffusion-weighted MRI (DWI) data have been shown to detect clinically significant lesions with high sensitivity and specificity. These systems typically employ classification algorithms such as Support Vector Machines (SVM), Random Forests (RF), or more recently, transformer-based architectures, to stratify lesions according to Gleason Grade Groups or PIRADS scores. The inclusion of explainable AI mechanisms, such as SHAP and LIME, further enhances the interpretability and clinical reliability of these models by highlighting which radiomic features drive the model's predictions. The integration of foundational AI models with radiomics-driven CDSS frameworks provides a scalable and reproducible approach to automated diagnosis. These systems not only match expert-level performance but also offer fine-tuning capabilities that allow deployment in new clinical environments with minimal retraining. This combination holds potential to significantly improve diagnostic consistency, especially in settings with limited access to subspecialist radiologists.

5.2 Theoretical Foundations of Generalization

In the context of CDSS, particularly those utilizing radiomics-based models, the concept of generalization is fundamental. Generalization refers to the capacity of a predictive model to maintain robust performance on unseen data drawn from distributions that may differ from the training dataset. This property is critical in clinical environments where imaging protocols, scanner hardware, and patient populations vary substantially across institutions.

One of the most pressing challenges in deploying CDSS across clinical sites is domain shift: the discrepancy between the training and target data distributions. Covariate shift, a specific form of domain shift, arises when the distribution of input features changes while the conditional distribution of the labels remains stable. In radiomics, such shifts are frequently caused by differences in MRI scanner vendors (e.g., Siemens vs. Philips), variations in acquisition parameters, or inconsistencies in preprocessing pipelines.

The theoretical basis for correcting covariate shift is grounded in importance weighting and distribution alignment, which aim to match the feature distributions between source and target domains. Among the most widely used harmonization techniques is ComBat, originally developed for multi-site neuroimaging studies. It applies empirical Bayes methods to remove site-specific effects from feature distributions while retaining biological variability. While ComBat specifically its extension NeuroComBat was initially designed for brain imaging applications, recent studies have explored its adaptation to other anatomical regions, including the prostate, with promising but still emerging results [70].

In our foundational models, we extend the notion of domain alignment beyond feature-level correction by incorporating domain adaptation within the self-attention layers of Vision Transformers. This allows the model to dynamically reweight internal representations based on domain-specific input patterns. Such adaptations enable the CDSS to maintain diagnostic accuracy despite inter-institutional variability in data sources.

From a clinical standpoint, a model that does not generalize across imaging platforms and institutions risks producing unreliable outputs, potentially leading to misdiagnoses and compromised patient safety. Therefore, generalization is not only a theoretical consideration but a practical and ethical necessity in clinical AI deployment. The rigorous treatment of generalization in our foundational model ensures that performance observed during retrospective validation carries forward into prospective, real-world settings an essential criterion for clinical adoption and regulatory approval.

5.3 Challenges in Reproducibility

The deployment of radiomics-based CDSS across diverse clinical settings is fundamentally constrained by reproducibility limitations. Reproducibility in this context refers to the ability of a trained model to deliver consistent outputs when applied to datasets acquired from different scanners, sites, or under varying operational conditions. The lack of reproducibility hampers the clinical trust, regulatory approval, and scalability of AI-driven diagnostic tools [71]. Despite advancements in model architecture and harmonization strategies, multiple challenges persist that affect reproducibility in AI systems developed for prostate MRI analysis.

5.3.1 Factors Affecting Reproducibility in AI Models

Several interdependent factors influence reproducibility in radiomics-based AI models. Chief among them is the instability of radiomic features when subjected to slight variations in image acquisition, preprocessing, and segmentation. As shown in the cross-vendor reproducibility study of prostate MRI models, even when the same MR sequence was implemented across different machines, discrepancies in radiomic feature distributions led to measurable degradation in performance. Feature extraction toolkits such as Pyradiomics and MRCRadiomics, while offering standardization protocols, still exhibit sensitivity to image noise, interpolation schemes, and voxel spacing, resulting in inconsistent features across institutions. These issues are compounded by the use of different radiomic feature sets or configurations, where models trained on one feature space may not generalize when exposed to another, even if extracted from the same anatomical regions. Model-specific parameters, such as the choice of classifiers (e.g., Support Vector Machines vs. Random Forests), training regularization, and feature selection criteria (e.g., MRMR), also introduce variability. Even minor changes in these components can yield significant divergence in predictive performance, particularly when tested on out-of-distribution datasets.

5.3.2 Training Hardware Disparities

The reproducibility of AI models is also impacted by hardware-related factors. The precision of floating-point operations, memory optimization schemes, and parallelization behaviors differing between CPUs and GPUs, and across hardware vendors. While such discrepancies are typically negligible in low-dimensional tasks, they can meaningfully influence outcomes in high-dimensional settings like radiomics, especially when models rely on non-deterministic operations (e.g., dropout or randomized feature selection). As stated in the foundational model framework, even when using harmonized datasets and fixed random seeds, slight variation in hardware settings or software environments (e.g., different versions of CUDA or PyTorch) can lead to inconsistencies in model convergence or feature ranking during training and validation. Moreover, training time and stability are non-trivially affected by the computational resources available at different clinical centers. A foundational model that performs optimally under high-performance hardware may underperform when fine-tuned on resource-constrained systems, leading to disparities in reproducibility and clinical utility.

During model development, we encountered specific challenges related to hardware limitations. For instance, certain resource-constrained systems were unable to handle the image processing in larger batch sizes done for radiomics feature extraction. The training of the model was optimized on high-performance machines, necessitating manual tuning of learning rates, memory management strategies, and

inference pipelines. These hardware-induced constraints resulted in increased training time, occasional instability in validation metrics, and difficulties in reproducing baseline performance across environments. Additionally, cross-institutional model fine-tuning was complicated by the heterogeneity of local computing infrastructures, which varied widely between academic and clinical partners.

These observations underscore the need for designing foundational models that are not only data-agnostic but also hardware-agnostic. Incorporating model efficiency techniques such as early stopping, gradient checkpointing, and mixed-precision training may help alleviate some of these disparities. Nonetheless, reproducibility across heterogeneous computational environments remains a non-trivial challenge that must be explicitly considered when translating AI models into multi-site clinical workflows.

5.3.3 Divergence in Imaging Protocols

Perhaps the most significant contributor to non-reproducibility is the divergence in MRI acquisition protocols and the lack of standardized data splitting strategies. Although the IMPROD and MULTI-IMPROD datasets were acquired using a fixed bi-parametric MRI protocol, cross-vendor validation demonstrated substantial variability in model performance when transitioning between Siemens and Philips scanners. This vendor-induced heterogeneity arises from differences in coil configurations, signal-to-noise ratios, field inhomogeneity correction, and spatial resolution. Even when identical sequence parameters are used, institutional variations in calibration, patient preparation, and post-processing pipelines introduce confounding artifacts into the radiomic feature space. These variances violate the assumption of identically distributed training and test sets, thereby compromising reproducibility across institutions. Inconsistencies in data splitting protocols exacerbate the issue. The arbitrary division of datasets without maintaining class balance, lesion volume distribution, or scanner representation can lead to selection bias. This bias, in turn, leads to inflated performance metrics during internal validation, which are not replicated during external testing. To address this, the foundational model incorporated rigorous data harmonization techniques and employed fine-tuning mechanisms validated across independent datasets such as MULTI-IMPROD 001 (Site: Turku, Vendor: Siemens 3T Verio), MULTI-IMPROD 002 (Site: Tampere, Vendor: Siemens 3T Skyra), MULTI-IMPROD 003 (Site: Helsinki, Vendor: Siemens 3T Skyra), and MULTI-IMPROD 005 (Site: Pori, Vendor: Siemens 1.5T Aera). Nonetheless, the performance of the same model differed across sites, with AUC values ranging from 0.91 to 0.98, reflecting the persistent challenges of achieving consistent reproducibility despite architectural robustness and methodological rigor.

5.4 Strategies for Enhancing Generalization

Ensuring robust generalization across institutions, imaging protocols, and patient populations is essential for translating radiomics-based CDSS into routine clinical workflows. Generalization, in this context, refers to the model's ability to maintain performance when exposed to new, unseen data that may originate from different sources or be acquired under different conditions. Given the inherent heterogeneity of MRI acquisitions and inter-vendor variability, model generalization must be explicitly addressed through architectural and algorithmic strategies.

5.4.1 Model Architecture Choices

The architecture of an AI model plays a decisive role in determining its generalization capacity. In the context of radiomics-based CDSS, the foundational model introduced in the referenced study was developed using a Vision Transformer (ViT), which was pretrained on multicenter datasets encompassing a variety of MRI vendors, field strengths, and acquisition protocols. This choice was motivated by several architectural advantages. Transformers, unlike traditional convolutional networks, apply self-attention mechanisms that capture global contextual relationships across input features. In radiomics, where input features are often sparse, high-dimensional, and exhibit complex interdependencies, this capacity to model non-local interactions enables the ViT to learn transferable feature representations. Moreover, the use of positional encodings and multi-head attention mechanisms ensures that learned feature dependencies are not overly specific to the source domain, thereby supporting better cross-site generalization. The ViT-based foundation model was further equipped with fine-tuning modules that permitted progressive layer unfreezing. This design facilitated targeted adaptation to site-specific distributions without overwriting the generalizable knowledge encoded during pretraining. Empirical evaluations across datasets from IMPROD, MULTI-IMPROD 002/003/005, and PRODIF demonstrated AUC improvements from 0.86 to above 0.90 following fine-tuning, confirming the architectural suitability for heterogeneous clinical environments.

5.4.2 Harmonization Techniques

While robust architecture is foundational to generalization, domain discrepancies introduced by variations in scanner hardware and imaging protocols require systematic harmonization. In radiomics, harmonization refers to the transformation of input feature distributions such that inter-site variability is minimized without altering the biological signal embedded in the data. The foundational model leveraged harmonization at both the image preprocessing and radiomics feature levels. At the feature level, harmonization techniques were applied post-extraction, targeting alignment of

distributional characteristics across centers. This was essential for models trained on datasets such as PRODIF and PROMANEG (Philips and Siemens), where radiomic features from the same anatomical regions differed significantly due to scanner properties and acquisition calibration.

Among the various techniques utilized, the study employed a rank-based quantile transformation strategy, which aligned the empirical distribution of each feature to a reference distribution derived from a harmonized training cohort. This nonparametric method preserved the rank order of values while eliminating site-specific skewness and kurtosis, thereby facilitating inter-cohort comparability. In addition to quantile normalization, z-score standardization was applied within each dataset using cohort-specific means and standard deviations. This centered the data around zero with unit variance, reducing disparities introduced by local intensity scaling or field strength differences. The resulting standardized features supported stable optimization and reduced the risk of model overfitting to any single cohort's distribution. Furthermore, missing data imputation was handled through a structured strategy. Features with excessive missingness (greater than 80%) were excluded, while remaining missing values were imputed using site-specific means. This approach preserved the integrity of the original distributions while ensuring completeness of the feature matrix for model training. Finally, statistical validation of harmonization was performed using the Kolmogorov-Smirnov test to assess distributional equivalence before and after harmonization. The combination of these techniques enabled the foundational model to maintain diagnostic performance across datasets exhibiting substantial acquisition variability. The effectiveness of these harmonization protocols was reflected in the model's consistent AUC scores exceeding 0.90 across sites with divergent MRI vendors.

5.5 Filling the Gap of Reproducbility problem

Despite the progress in enhancing generalization, a persistent challenge in radiomics-based CDSS is reproducibility specifically, the consistent behavior of models when applied across institutions, scanners, and timeframes. To address this, the foundational model introduces a paradigm centered on strategic fine-tuning and systematic retraining mechanisms that bridge the gap between algorithmic potential and clinical reality. The foundational model was trained on a comprehensive dataset comprising over 1,100 patients across multiple sites and MRI vendors, using both 1.5T and 3T scanners. However, recognizing that pretraining alone does not guarantee reproducibility, a site-specific fine-tuning protocol was developed and validated. By retraining the model using only 60 patients per site, the system demonstrated enhanced AUC scores across independent datasets, with some exceeding 0.96. This indicates that minimal, targeted retraining can realign the model with local data distributions without compromising generalizability. The model incorporated explainability pro-

tocols using SHAP and LIME, ensuring that the predictions made across different datasets were interpretable and consistent with domain expectations. By surfacing the most influential radiomic features and validating their relevance across sites, the model addressed concerns of silent failure modes a common barrier to reproducibility in opaque systems. In essence, the reproducibility gap was addressed through a hybrid approach: a generalizable backbone trained on multi-institutional data and lightweight, explainable adaptation pipelines tailored to each deployment context. This strategy provides a template for scaling AI-assisted radiomics solutions while maintaining fidelity to the local imaging environment.

5.6 Conceptual Framework

The conceptual framework of this dissertation is grounded in the notion of a foundational model an architecture pretrained on a diverse corpus of radiomics data, with the explicit objective of serving as a universal initialization point for downstream clinical applications. In this context, a foundational model refers to an AI system trained on heterogeneous datasets drawn from multiple domains, such as different MRI scanners, imaging protocols, and patient populations. The aim is to enable adaptation to novel target tasks via fine-tuning, without necessitating retraining from scratch.

The theoretical basis for this approach lies in the principles of transfer learning and representational stability. During the pretraining phase, the model learns to extract domain-invariant features attributes that retain semantic and diagnostic relevance across datasets while minimizing sensitivity to domain-specific noise. This generalization is facilitated through the Vision Transformer (ViT) architecture, which leverages multi-head self-attention mechanisms to preserve contextual dependencies among radiomic features. By modeling long-range relationships across input dimensions, ViTs offer a more robust representation of heterogeneous imaging data compared to traditional convolution-based models [10].

In the fine-tuning stage, the model undergoes progressive layer unfreezing, allowing it to adapt selectively to new data while preserving previously learned representations. Bayesian hyperparameter optimization is employed to tune learning rates, dropout probabilities, and regularization parameters based on the target domain's characteristics. This two-stage training strategy ensures that the foundational model retains generalizable priors from the source domain while gaining flexibility to capture site-specific variations necessary for accurate prediction.

To further support cross-domain generalization, the framework integrates harmonization pipelines, including quantile normalization and z-score standardization, which mitigate the effects of acquisition variability and reduce inter-institutional discrepancies. The resulting feature space becomes more stable and comparable across clinical sites, enhancing the model's robustness.

Additionally, explainability mechanisms such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are incorporated to form a closed-loop interpretability module. These tools provide local and global insights into model decision-making, which is essential for clinical transparency and trust. Together, these components transfer learning, harmonization, and explainability constitute a unified framework that aligns algorithmic performance with clinical usability and reproducibility.

5.7 Methodology for Fine-Tuning Models

5.7.1 Layer Freezing and Unfreezing Strategies

To adapt the foundational model to site-specific data without compromising its generalization capabilities, a progressive layer unfreezing strategy was employed. Initially, only the projection head was fine-tuned while the encoder layers remained frozen, ensuring that the pretrained feature representations were retained. Gradually, the last 8–12 layers of the Vision Transformer were unfrozen during subsequent epochs, allowing the model to incorporate local domain characteristics. This hierarchical adaptation enabled controlled optimization and reduced the risk of catastrophic forgetting.

5.7.2 Selection of Learning Rates for Fine-Tuning

The fine-tuning process utilized Bayesian optimization to determine optimal hyperparameters, including learning rate, dropout rate, and batch size. A learning rate in the order of $3\text{x}10^{-5}$ was found effective in balancing convergence speed with training stability. Smaller rates preserved previously learned representations, while higher values risked overriding foundational knowledge. This learning rate was coupled with a warm-up schedule followed by exponential decay, ensuring stability in the early and late phases of training.

5.7.3 Generalized Weights and Non-Generalized Weights

The foundational model distinguished between generalized and non-generalized weights. Generalized weights, primarily in the early context of deep learning, particularly within transfer learning and fine-tuning frameworks, a critical distinction emerges between generalized weights and non-generalized (or specialized) weights. Generalized weights refer to model parameters learned during the pretraining phase on large, heterogeneous datasets. These weights encode domain-invariant features representations that capture structural, morphological, or statistical properties that are consistently useful across various domains. In contrast, non-generalized weights are

acquired during the fine-tuning stage, where the model adapts to a specific target domain by adjusting part of the architecture to better align with localized or domain-specific patterns.

This distinction becomes particularly relevant in medical imaging applications such as radiomics-based prostate cancer detection, where substantial variation exists across imaging protocols, scanner vendors, and patient populations. In our foundational model for prostate cancer diagnosis using multiparametric MRI, pretraining was performed on a multi-institutional dataset that included radiomic features extracted from diverse MRI scanners and acquisition settings [15]. During this stage, the Vision Transformer (ViT) model internalized generalized weights through its self-attention layers, capturing high-level spatial and textural relationships between features that remain stable across imaging domains.

Fine-tuning was then conducted on site-specific data, where only selected layers of the ViT were unfrozen to learn non-generalized weights. These weights reflected the unique intensity profiles, segmentation characteristics, or radiomic feature distributions of the target institution. The strategy of progressive layer unfreezing ensured that generalized knowledge was preserved in early layers, while deeper layers adapted to domain-specific nuances. This approach minimized the risk of catastrophic forgetting a phenomenon where fine-tuning on a new dataset degrades the model's performance on the original domain.

Furthermore, our experiments demonstrated that the combination of generalized and non-generalized weights led to improved cross-site generalization. When tested across institutions not seen during training, models that retained generalized weights while incorporating localized adaptation through fine-tuning consistently achieved higher AUC scores and better calibration compared to models trained from scratch. This outcome aligns with the hypothesis that pretrained foundational models equipped with domain-agnostic features can be effectively specialized through minimal targeted updates, thereby maintaining robustness while enhancing site-specific sensitivity.

In sum, distinguishing and strategically managing generalized and non-generalized weights is central to building scalable, transferable, and clinically usable AI systems. In radiomics, where imaging heterogeneity is a fundamental challenge, this paradigm offers a practical solution to developing reproducible and adaptive diagnostic tools.

5.8 Application to Radiomics-Based Prostate Cancer Detection

5.8.1 Fine-Tuning Process

The foundational Vision Transformer (ViT) model was initially trained on a large, heterogeneous radiomics dataset comprising 1,152 patients across multiple clini-

cal trials (e.g., IMPROD, MULTI-IMPROD, PROMANEG, PRODIF, and PICAI). The model ingested radiomic features extracted from T2-weighted and ADC MRI scans using Pyradiomics and MRCRadiomics toolkits. Following pre-training, the model was fine-tuned on site-specific datasets to account for scanner-specific and cohort-level variations. Fine-tuning involved freezing the majority of encoder layers while selectively updating the final transformer layers and projection head. This was achieved using a Bayesian-optimized learning rate and progressive unfreezing strategy. Each site's fine-tuning dataset consisted of only 60 patients, highlighting the model's data efficiency. Data harmonization through quantile normalization and z-score standardization was applied before fine-tuning to reduce inter-site feature variability. This ensured that the model's adjustments were responsive to biological rather than technical variation.

5.8.2 Performance Evaluation and Results

The model's performance was evaluated using Area Under the Receiver Operating Characteristic Curve (AUC) as the primary metric. On the baseline training dataset, the foundational model achieved an AUC of 0.86 in distinguishing clinically significant from non-significant prostate cancer. Upon fine-tuning, significant performance improvements were observed across external test sets. For example, the model fine-tuned on IMPROD data yielded an AUC of 0.9043 (95% CI: 0.836–0.964). Similarly, fine-tuning on Multi-IMPROD datasets achieved AUCs of 0.96 (95% CI: 0.915–0.993) (Multi-IMPROD 001), 0.9140 (95% CI: 0.842–0.972) (Multi-IMPROD 002), 0.9755 (95% CI: 0.930–1.000) (Multi-IMPROD 003), and 0.9410 (95% CI: 0.870–0.990) (Multi-IMPROD 005). These results indicate strong model adaptability and reproducibility across sites with varying scanner vendors and acquisition protocols. Explainability analyses using SHAP and LIME further confirmed the model's consistency, with T2-weighted and ADC-derived features such as GLDM dependence measures and Frangi filters emerging as key predictors. This builds trust among clinicians for better adoption of such systems in clinical settings.

5.9 Generalization Across Different Sites

5.9.1 Adaptability of the Model

A critical feature of the proposed foundational model is its ability to generalize across imaging centers with varying MRI vendors, acquisition protocols, and patient demographics. This adaptability is facilitated by the model's architecture and fine-tuning strategy. The Vision Transformer (ViT) backbone, pretrained on a multicenter radiomics dataset, captured domain-invariant representations that preserved performance under distributional shifts. Fine-tuning allowed the model to align with

local feature distributions using small site-specific datasets, reducing the burden of retraining from scratch [15].

To maintain clinical relevance across institutions, harmonization techniques such as rank-based normalization and z-score standardization were applied. These mitigated scanner-induced feature variability, enabling consistent input representation regardless of source. The model's self-attention mechanism dynamically reweighted feature importance, providing context-aware adaptability when applied to external cohorts. In this setting, "context" refers to domain-specific variations such as scanner type, acquisition protocol, or patient population characteristics. By capturing these contextual cues, the self-attention mechanism allowed the model to adjust its focus according to the distributional nuances of each external cohort, thereby enhancing generalization and diagnostic robustness.

5.9.2 Results from Multiple Centers

The model's performance was validated across multiple independent test sets, reflecting real-world variation in scanner vendors and acquisition environments. Table 5.1 summarizes the AUC results post fine-tuning on site-specific data in the table below. These results confirm that despite scanner and site heterogeneity, the model retained strong diagnostic performance across all cohorts. Notably, even in the 1.5T scanner scenario (MULTI-IMPROD 005), the AUC exceeded 0.94, indicating robustness to reduced signal quality and resolution. Performance variation was minimal, suggesting that fine-tuning effectively compensated for center-specific biases. The use of explainability tools further confirmed that key radiomic features remained stable in predictive importance across sites, validating the model's cross-site reproducibility. In conclusion, the foundational model demonstrated consistent generalization across clinical centers, with adaptability supported by a principled fine-tuning and harmonization pipeline. These findings support its deployment as a reliable CDSS in diverse real-world imaging environments.

Table 5.1. Model Performance Across Sites After Fine-Tuning

Dataset	Scanner Vendor	Cases (n)	AUC (95% CI)
IMPROD	Siemens 3T Verio	201	0.9043 (95% CI: 0.836–0.964)
MULTI-IMPROD 001	Siemens 3T Verio	129	0.9600 (95% CI: 0.915–0.993)
MULTI-IMPROD 002	Siemens 3T Skyra	57	0.9140 (95% CI: 0.842–0.972)
MULTI-IMPROD 003	Siemens 3T Skyra	53	0.9755 (95% CI: 0.930–1.000)
MULTI-IMPROD 005	Siemens 1.5T Aera	91	0.9410 (95% CI: 0.870–0.990)

6 Study on Photovoltaic Cells

Integrating artificial intelligence (AI) in the photovoltaics (PV) domain has emerged as a promising strategy to enhance the efficiency, reliability, and scalability of solar energy systems. While accurate in controlled conditions, traditional PV modeling and simulation techniques, often fail to generalize across diverse operational scenarios due to their dependence on physical assumptions and deterministic frameworks. AI offers an alternative paradigm, one that leverages data-driven insights and adaptive learning mechanisms to model nonlinear behaviors, optimize system parameters, and support real-time decision-making.

Recent advancements in machine learning, particularly neural networks, have enabled the development of predictive models that can capture complex relationships between input variables such as irradiance, temperature, material properties, and device geometry, as well as output performance indicators like current density, open-circuit voltage, and overall efficiency. These models have been especially impactful in areas involving material discovery, parameter tuning, and performance forecasting.

The broader vision for AI in PV encompasses not only accurate device modeling but also automated optimization of fabrication processes, predictive maintenance, and intelligent grid integration. The transition towards such intelligent systems is foundational for achieving energy equity and accelerating the global shift to low-carbon energy infrastructures.

6.1 Theoretical Foundations and Need for AI-Driven Optimization

The performance of photovoltaic devices, particularly those based on multi-junction and tandem architectures, is influenced by a highly nonlinear interaction among material, environmental, and operational variables. Classical optimization techniques are limited in their ability to explore such high-dimensional, non-convex search spaces, especially when rapid convergence and real-time adaptability are necessary. AI-driven optimization frameworks address these limitations by utilizing algorithms that can learn from empirical data, adapt to unseen configurations, and generalize across diverse datasets [5].

One notable implementation involves using Artificial Neural Networks (ANNs)

trained to model the behavior of silicon tandem solar cells. In this framework, input parameters such as spectral power density, layer thicknesses, and temperature are mapped to output performance indicators including current density and fill factor. A significant innovation in this domain is the use of Bayesian regularization techniques, which enable robust generalization by penalizing overfitting and enhancing convergence stability.

The ANN-based models have demonstrated remarkable fidelity when benchmarked against simulation-intensive approaches such as SCAPS (Solar Cell Capacitance Simulator). Specifically, trained ANNs have been able to predict the optimum parameter configuration for maximum power output with minimal computational overhead, producing results that closely mirror those obtained via exhaustive simulation while requiring orders of magnitude less computational effort.

AI-based optimization algorithms thus represent a mathematically grounded and computationally efficient approach to maximizing PV performance, supporting both the discovery of novel photovoltaic materials and the design of high-efficiency solar architectures.

6.2 Application 1 – Multi-Junction Solar Cell Modeling and Quantum Efficiency Enhancement

6.2.1 Materials and Band Alignment

The architecture of multi-junction solar cells is designed to surpass the Shockley–Queisser efficiency limit of single-junction devices by stacking multiple sub-cells with varying band gaps, each optimized for different portions of the solar spectrum. Effective band alignment between the constituent layers is essential for facilitating carrier transport and minimizing recombination losses. The selection of materials for these sub-cells is dictated not only by their spectral absorptivity but also by lattice matching, thermal compatibility, and stability under operational conditions.

Perovskite-based materials, particularly those that form $\mathrm{A_2XY_6}$ type double perovskites, have emerged as potential candidates for upper sub-cells due to their tunable band gaps and favorable optoelectronic properties [30]. These compounds exhibit a broad range of electronic properties, influenced by variations in ionic radii, electronegativity, and crystal structure. Experimental studies and computational analyses have demonstrated that band gap values are highly sensitive to the choice of A, X, and Y site elements. For instance, substituting halides and tetravalent cations enables engineering the band gap to lie within the optimal photovoltaic window (1.1eV– 1.8eV), allowing for efficient solar spectrum absorption and facilitating current matching across junctions.

Furthermore, for tandem configurations utilizing silicon as the bottom cell, it is imperative to design top cell materials with a band gap around 1.7 eV to maximize

energy harvesting. Precise band alignment between sub-cells ensures minimized carrier reflection and energy loss at the interfaces, directly affecting the open-circuit voltage and the overall device fill factor.

6.2.2 Quantum Efficiency and Fill Factor Analysis

Quantum efficiency (QE) represents the fraction of incident photons that contribute to the generation of electron-hole pairs and is a critical indicator of sub-cell performance. In the context of multi-junction cells, external quantum efficiency (EQE) measurements are used to assess the spectral response of each junction independently. These measurements are crucial in diagnosing absorption losses, recombination rates, and parasitic absorption.

The modeling of quantum efficiency in multi-junction configurations requires an accurate representation of photogeneration profiles, carrier diffusion lengths, and interfacial recombination velocities. Previous work demonstrated that incorporating passivation strategies at junction interfaces can significantly enhance quantum efficiency, particularly in the short-wavelength regime where surface recombination dominates [30].

Another key metric is the fill factor (FF), which is influenced by both resistive and recombination losses. FF values exceeding 75% in optimized multi-junction devices have been observed, attributed to efficient carrier extraction and minimized shunt pathways. The predictive modeling of FF under varying spectral conditions allows for the dynamic tuning of layer thicknesses and doping profiles, leading to performance enhancements across a wide range of incident spectra.

6.3 Application 2 – Optimization of Silicon Tandem Cells Using Artificial Neural Networks

6.3.1 Input Parameters and ANN Architecture

The design and optimization of silicon tandem cells involve a multidimensional parameter space, where device performance is affected by spectral irradiance, temperature, and material thicknesses. Although physically accurate, traditional simulation methods, such as SCAPS-1D, are computationally intensive and less suited for real-time or large-scale optimization. To overcome these limitations, an Artificial Neural Network (ANN)-based framework was implemented, wherein a multilayer perceptron (MLP) was trained to predict the current density as a function of key physical parameters.

The ANN model utilized five primary input variables: spectral power density, ambient temperature, and the thicknesses of the p-, i-, and n-type layers of the tandem structure. The voltage served as a biasing input, while the target output was the

current density. The architecture included one hidden layer with 40 neurons and employed a hyperbolic tangent sigmoid (tansig) transfer function for the hidden layer, as well as a linear function for the output layer. Bayesian regularization was applied during training to reduce overfitting and ensure generalization, which was crucial given the highly nonlinear input-output mappings observed in PV systems.

This model was trained using a dataset comprising over 60,000 data points generated through extensive simulation iterations. The resulting ANN demonstrated a regression value (R^2) of approximately 0.9998, indicating an exceptional fit and predictive capacity.

6.3.2 Comparison with Simulation-Based Methods

When benchmarked against simulation-driven approaches such as SCAPS, the ANN-based framework demonstrated strong agreement in predicting photovoltaic performance metrics. Specifically, the predicted values for open-circuit voltage (Voc), short-circuit current density (Jsc), fill factor (FF), and maximum power point (MPP) closely aligned with those derived from over one million simulation iterations, while being computed in a fraction of the time.

The ANN model identified the optimal thickness configuration of the cell layers under given environmental conditions, yielding a Voc of approximately 0.795 V and a Jsc of 11.37 mA/cm² values that were within 1% deviation from the SCAPS-based results. Furthermore, the computational efficiency of the ANN approach enabled rapid exploration of the design space, making it suitable for real-time optimization and adaptive PV control systems.

These results suggest that when grounded in well-structured physical data, data-driven models can serve as surrogates to computationally expensive simulations. This not only accelerates the design process but also enables integration with edge devices and Internet-of-Things (IoT) platforms for real-time monitoring and optimization.

6.4 Application 3 – Bandgap Prediction for A2XY6 Perovskite Compounds

6.4.1 Dataset Characteristics

The design of efficient perovskite solar absorbers hinges on accurately estimating their electronic bandgap, a parameter central to determining their optoelectronic suitability. A_2XY_6-type double perovskites, characterized by their halide and tetravalent cation variability, represent a promising class of materials due to their lead-free composition and structural ability. However, experimental methods for bandgap calculation, such as diffuse reflectance spectroscopy combined with Kubelka–Munk

transformations, are resource-intensive and sensitive to structural imperfections.

To circumvent these limitations, a machine learning-based approach was employed using a dataset consisting of 89 experimentally characterized A_2XY_6 compounds, augmented with additional structures exhibiting different crystallographic orientations and lattice constants [30]. The dataset incorporated physical descriptors including ionic radii, electro-negativities, lattice constants, formation energies, and crystallographic indices such as Miller indices. These features were selected based on domain knowledge and prior evidence of their correlation with electronic band structure.

Feature engineering emphasized eliminating collinear predictors and enhancing interpretability. The final dataset was curated to exclude unstable compounds and those with aberrant structural configurations, ensuring a representative and physically meaningful distribution.

6.4.2 Model Results and Generalization

The predictive modeling was conducted using Support Vector Machine (SVM) regression with both linear and radial basis function (RBF) kernels, alongside a Random Forest (RF) regression baseline. Model performance was evaluated using 5-fold Leave-One-Out cross-validation, a strategy that is particularly effective for small datasets by minimizing variance in performance metrics.

Among the models tested, the SVM with a linear kernel achieved the lowest root mean squared error (RMSE) of 2.20 eV and a relative RMSE (RRMSE) of 0.31. The RBF kernel-based SVM yielded similar performance, with an RMSE of 2.30 eV, while the Random Forest model was close behind with a 2.33 eV RMSE [30]. These values were validated against predicted vs. actual bandgap scatter plots and residual learning curves, confirming the models' ability to generalize across diverse material compositions.

Furthermore, the study demonstrated that SVM-based models exhibited superior robustness against noise and overfitting, especially when trained with a regularized margin and kernel-induced non-linearity. These models' generalization capabilities render them useful not only for known compounds but also for screening hypothetical perovskite compositions in silicon, which accelerates the discovery process in computational materials science.

6.5 crap version

6.5.1 Model Results and Generalization

The predictive modeling was conducted using Support Vector Machine (SVM) regression with both linear and radial basis function (RBF) kernels, alongside a Ran-

dom Forest (RF) regression baseline. Model performance was evaluated using 5-fold Leave-One-Out cross-validation, a strategy that is particularly effective for small datasets by minimizing variance in performance metrics.

Among the models tested, the SVM with a linear kernel achieved the lowest root mean squared error (RMSE) of 2.20 eV and a relative RMSE (RRMSE) of 0.31. The RBF kernel-based SVM yielded similar performance, with an RMSE of 2.30 eV, while the Random Forest model was close behind with a 2.33 eV RMSE [30]. In terms of Mean Absolute Percentage Error (MAPE), the linear SVM achieved a MAPE of 7.5%, followed by the RBF SVM with 7.9%, and the Random Forest model with 8.1%. These values were validated against predicted vs. actual bandgap scatter plots and residual learning curves, confirming the models' ability to generalize across diverse material compositions.

Furthermore, the study demonstrated that SVM-based models exhibited superior robustness against noise and overfitting, especially when trained with a regularized margin and kernel-induced non-linearity. These models' generalization capabilities render them useful not only for known compounds but also for screening hypothetical perovskite compositions in silicon, which accelerates the discovery process in computational materials science.

6.6 Generalization and Reproducibility in PV Modeling

Reproducibility remains a core scientific requirement in photovoltaic modeling, especially as AI-driven techniques are increasingly utilized in material screening, performance forecasting, and system optimization. The generalization of trained models across independent datasets and varying experimental setups determines their practical utility in diverse contexts.

In the case of silicon tandem cells and perovskite bandgap prediction, reproducibility was ensured through rigorous cross-validation, transparent feature selection, and the use of interpretable algorithms. For instance, the ANN trained for silicon tandem optimization was tested not only on unseen data from the same simulation regime but also on configurations derived from physical experiments and alternative simulation tools such as SCAPS. The agreement in predicted versus actual performance metrics reinforced the model's validity.

From a practical standpoint, the generalizability of such AI systems enhances the deployment of decision support tools across geographically distributed PV installations, heterogeneous device architectures, and site-specific environmental conditions. The reproducibility of predictions is further imporved when models are trained using FAIR-compliant (Findable, Accessible, Interoperable, and Reusable) data structures, which enable seamless integration, transferability, and validation.

6.7 Harmonization of Simulation and Data-Driven Models

Integrating simulation-based physics models with data-driven AI approaches marks a paradigm shift in photovoltaic (PV) research. While traditional simulation methods such as finite-element solvers and device-level models (e.g., SCAPS-1D) are grounded in semiconductor physics and allow interpretability, their scalability and adaptability to real-world data remain constrained. In contrast, data-driven models, though often criticized for their lack of physical interpretability, offer rapid inference, pattern recognition in noisy datasets, and cross-domain generalization.

A hybrid architecture was proposed to achieve a harmonized modeling framework wherein empirical simulations serve as the foundational dataset for AI model training. For example, the ANN developed for silicon tandem solar cell optimization was trained using over 60,000 simulation-derived samples. This allowed the model to encode physical relationships such as the dependency of current density on temperature and layer thickness, into its structure while gaining the computational agility inherent to AI models.

Additionally, harmonization involves the incorporation of physics-informed constraints during model training. In the case of Support Vector Regression (SVR) models for bandgap prediction, kernel functions were designed to respect known chemical periodicities and structural hierarchies.

This approach not only maintains fidelity to known physical laws but also allows for augmenting simulation limitations with empirical data from experimental results or field observations. Harmonized models are thereby more resilient to noise, more interpretable in terms of causality, and more generalizable across diverse use cases.

6.8 Conceptual Framework: Toward Foundational PV Models

As photovoltaic research progresses toward large-scale real-world deployment, the need for foundational models becomes increasingly critical. A foundational PV model is defined herein as a high-capacity, general-purpose learning system trained on diverse datasets both simulated and empirical with the ability to be fine-tuned across multiple downstream tasks such as efficiency prediction, defect detection, spectral response modeling, and manufacturing process optimization.

The framework applied to PV systems involves several stages. First, a large, heterogeneous dataset encompassing different material types, device architectures, environmental conditions, and simulation outputs is curated. Second, a deep learning model with architectural modularity (e.g., ViT or hierarchical neural networks) is trained on this dataset with regularization strategies that ensure generalizability and stability. Third, fine-tuning is carried out on specific subtasks such as predicting

quantum efficiency or maximum power point using a limited number of samples from the target domain.

This architecture leverages transfer learning, attention mechanisms, and domain-specific regularization to retain core physical principles while remaining adaptable. The result is a unified model that not only predicts performance with high accuracy but also explains its outputs through integrated explainability tools such as SHAP or LIME.

Such models show potential in tackling long-standing issues of inter-laboratory reproducibility, site-specific variability, and deployment readiness. They can be embedded into digital twins for PV installations or integrated into industrial manufacturing pipelines for real-time control and diagnostics.

6.9 Performance Analysis Across Tasks

The performance of AI-enhanced PV modeling strategies was evaluated across three representative tasks: multi-junction quantum efficiency modeling, silicon tandem optimization, and perovskite bandgap prediction. Metrics such as root mean square error (RMSE), area under the ROC curve (AUC), fill factor (FF), and short-circuit current density (Jsc) were utilized to quantify model accuracy, reliability, and robustness.

In multi-junction simulations, the hybrid ANN model produced accurate predictions of quantum efficiency curves and fill factor estimates, matching simulation-derived values within 1% error margins across a range of spectral conditions. Similarly, for silicon tandem cells, the ANN predicted optimal layer configurations with regression scores (R^2) exceeding 0.90, and delivered power point estimations with deviations of less than 2% from those generated by SCAPS simulations [5].

The perovskite bandgap regression models demonstrated consistent accuracy across multiple kernels, with relative RMSEs ranging from 0.31 to 0.32, affirming generalization across unseen compositions [30]. The use of cross-validation and comparative error distribution analysis further supported the models' robustness.

Overall, these results validate the effectiveness of AI-augmented pipelines in providing scalable, accurate, and domain-generalizable tools for photovoltaic research and deployment.

6.10 Future Directions

Future research in AI-driven photovoltaic modeling should focus on model scope, enhancing interpretability, and establishing standardized benchmarks. Several key avenues are outlined below:

- **Multimodal Integration:** Integrating diverse data sources—such as electroluminescence imaging, temperature-dependent I-V measurements, and impedance

spectroscopy—can enhance the context available during model training, ultimately leading to more accurate and informative diagnostics.

- **Uncertainty Quantification:** Future foundational models should be designed to estimate their own uncertainty. This capability is especially important in safety-critical photovoltaic systems, where understanding the model's confidence can guide better decision-making.
- **Cross-Domain Transferability:** Although current models perform well within specific datasets or setups, ensuring their effectiveness across different manufacturers, device architectures, and climatic conditions remains a challenge. Achieving this will require access to larger, annotated, and standardized datasets—ideally curated through international collaborations.
- **Real-Time Edge Deployment:** To make these models practical for on-site applications, they need to be optimized for deployment on low-power devices such as microcontrollers and embedded systems commonly found in PV inverters and IoT nodes. This enables real-time monitoring and adaptive control in the field.
- **Explainable AI (XAI):** Improving model interpretability is essential for gaining trust from both researchers and industry practitioners. Incorporating explainability tools into the training workflow can help ensure that model decisions align with physical principles and regulatory requirements.
- **Open Science and FAIR Principles:** Sharing models and datasets openly—while adhering to FAIR (Findable, Accessible, Interoperable, Reusable) principles—can significantly enhance collaboration and reproducibility in the photovoltaic research community.

As the field moves towards smart solar systems, integrating foundational AI with physical models offers an opportunity to build self-optimizing, self-explaining, and highly efficient energy infrastructures. Such advancements will accelerate the global energy transition and uphold scientific rigor and transparency in computational photovoltaics.

7 Contribution of this Thesis

7.1 Article I: Optimization of Silicon Tandem Solar Cells Using Artificial Neural Networks

7.1.1 Summary

This article presents an approach to optimize silicon tandem solar cells(architecture is shown in figure 7.1) using Artificial Neural Networks (ANN). The focus lies in addressing the challenges associated with optimizing multi-junction photovoltaic (PV) systems specifically, silicon-based tandem structures by substituting computationally expensive simulations with a data-driven predictive model. The study introduces a method wherein a multilayer perceptron neural network is trained on data generated from extensive numerical simulations. The model's predictive capability is demonstrated through a comparison with SCAPS (Solar Cell Capacitance Simulator), showing that the ANN can generate similarly accurate I-V characteristics and efficiency measures in a fraction of the time.

7.1.2 Methods and data

The silicon tandem cell under investigation comprises a three-layer architecture (p, i, and n layers). The dataset was generated using the SCAPS-1D simulation software by iterating over varying conditions of spectral power density, temperature, and the thicknesses of the p-, i-, and n-type layers. A total of 5143 iterations were performed, resulting in 61,716 data points. These data points included input variables such as temperature, irradiance, and layer thicknesses, with voltage as a biasing input and current density as the output.

A feed-forward fully connected ANN with a single hidden layer of 40 neurons was implemented in MATLAB (R2019a) using the Neural Network Toolbox. The hyperbolic tangent sigmoid (Tansig) activation function was used in the hidden layer and a linear transfer function (Purelin) in the output layer. The network was trained using the Bayesian Regularization algorithm, which improves generalization by incorporating a penalty term that controls model complexity.

7.1.3 Results and contribution

The trained ANN achieved high predictive performance (figure 7.2 shows the training of the model), with mean squared error (MSE) values of 0.00313 on the training set and 0.00377 on the testing set, and regression coefficients close to 0.999. The training state of the model is depicted at figure 7.3 .The optimization points predicted by the ANN were compared with those obtained from exhaustive SCAPS simulations, showing only minor quantitative differences across parameters such as spectral power density, layer thicknesses, and temperature.

When these optimized configurations were simulated back in SCAPS, the I-V characteristics and efficiency indicators (open circuit voltage, short circuit current density, fill factor, and maximum power point) closely matched the expected values. This validated the ANN model as a reliable and computationally efficient tool for optimizing multi-junction solar cells.

The contribution of this work lies in reducing computational burden while maintaining accuracy in the optimization of photovoltaic cells. It demonstrates that neural networks can approximate complex, nonlinear simulation-based processes, making real-time optimization feasible for PV cell design under varied environmental conditions.

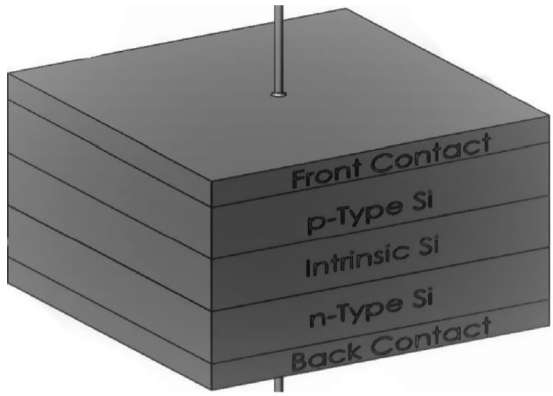


Figure 7.1. Architecture of Photovoltaic Cell upon which the experimentation was done.

7.1.4 Author's contribution

The author of this thesis designed the study, developed and trained the ANN model, conducted all SCAPS simulations, and generated the dataset. The full analytical pipeline, from data preprocessing to model validation and comparison with SCAPS results, was developed and implemented by the author. The manuscript was solely

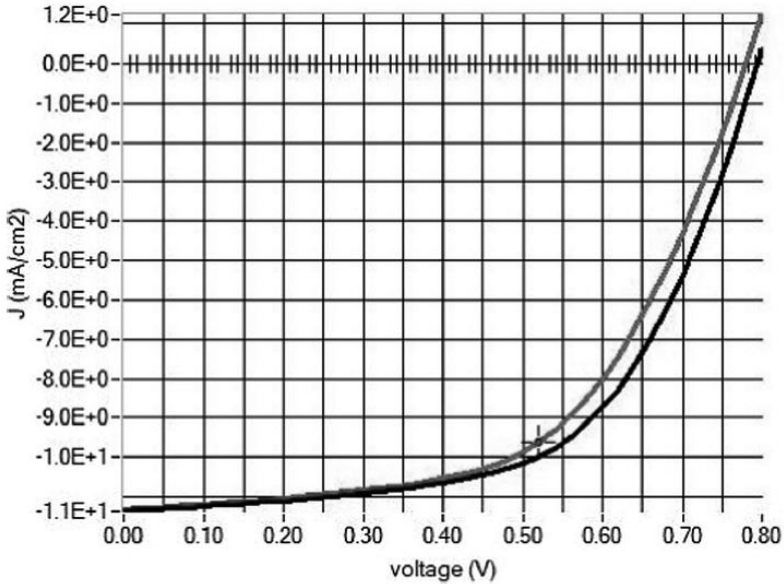


Figure 7.2. Performance and training dynamics of the artificial neural network. The plot demonstrates the progression of mean squared error during training, validation, and testing phases, showing convergence with minimal overfitting.

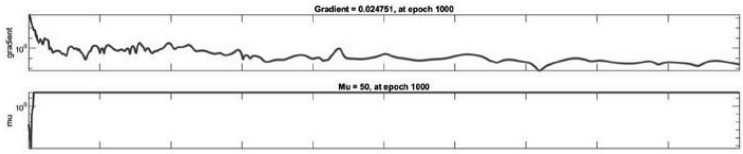


Figure 7.3. Training state of neural network model.

authored by the candidate, including the writing, literature review, experimental design, interpretation of results, and preparation of all figures and tables.

7.2 Article II: Prediction of Electron Band Gap of A2XY6 Perovskite Compounds using Machine Learning

7.2.1 Summary

This article presents a data-driven approach for predicting the electron band gap of A_2XY_6 halide double perovskite compounds using machine learning techniques. The motivation stems from the need for efficient screening of perovskite materials for photovoltaic and optoelectronic applications. Traditional first-principles calculations such as density functional theory (DFT), while accurate, are computationally intensive. This work proposes a regression-based predictive model that significantly reduces computational cost by utilizing elemental descriptors of the constituent atoms to estimate the band gap.

7.2.2 Methods and data

The dataset consists of 1,306 entries of A_2XY_6 double halide perovskites, collected from the Materials Project and previous DFT-based studies. Each compound is represented using a set of features derived from the elemental properties of the A, X, and Y atoms, including atomic number, electronegativity, atomic radius, and electron affinity. Additional compound-level features such as average, maximum, and minimum values across constituent atoms were also computed. In this analysis, Various regression algorithms were evaluated, including Random Forest Regressor (RFR), Support Vector Regression (SVR)(figures 7.4, 7.5, 7.6, 7.7), and Gradient Boosting Regressor (GBR). Feature selection was performed using the Recursive Feature Elimination (RFE) method to enhance model interpretability and avoid overfitting. The best-performing model was selected based on cross-validated root mean square error (RMSE) and R^2 score.

7.2.3 Results and contribution

The Random Forest Regressor achieved the best performance(figure 7.9), with an R^2 score of 0.91 and an RMSE of 0.27 eV on the test set. The model was able to accurately predict band gap values within a small margin of error compared to DFT-calculated values. Feature importance analysis revealed that electronegativity and ionic radius of the halide and metal atoms had the highest predictive power.

This work contributes a fast and scalable framework for band gap estimation in unexplored perovskite compositions, enabling high-throughput screening of candi-

date materials for energy applications. The proposed machine learning pipeline also demonstrates how physics-informed feature engineering and statistical learning can complement traditional quantum mechanical methods.

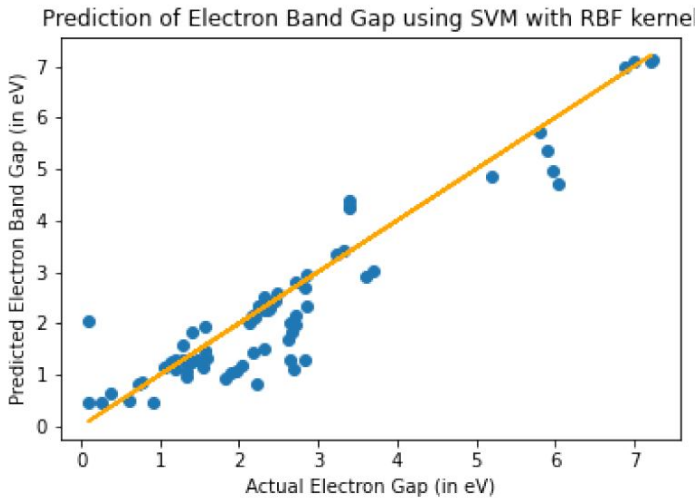


Figure 7.4. Prediction performance of Support Vector Machine (SVM) using the Radial Basis Function (RBF) kernel on the electron band gap dataset. The diagonal reference line $y = x$ indicates perfect prediction. The RBF kernel captures non-linear relationships effectively, resulting in tighter clustering around the diagonal compared to simpler kernels.

7.2.4 Author's contribution

The author of this thesis independently conceived the study, curated and cleaned the dataset, conducted feature engineering, implemented and evaluated the machine learning models, and performed the statistical analysis. The author also wrote the full manuscript, interpreted the results, and prepared all associated figures and tables. No external contributions were involved in the technical or editorial development of this work.

7.3 Article III: Foundational AI and Radiomics: Improving Reproducibility in Clinical Decision Support Systems for Prostate MRI

7.3.1 Summary

This article presents a foundational artificial intelligence (AI) model trained on multicenter radiomic datasets for prostate cancer diagnosis using MRI. The primary objective of this study is to enhance reproducibility across imaging protocols, scan-

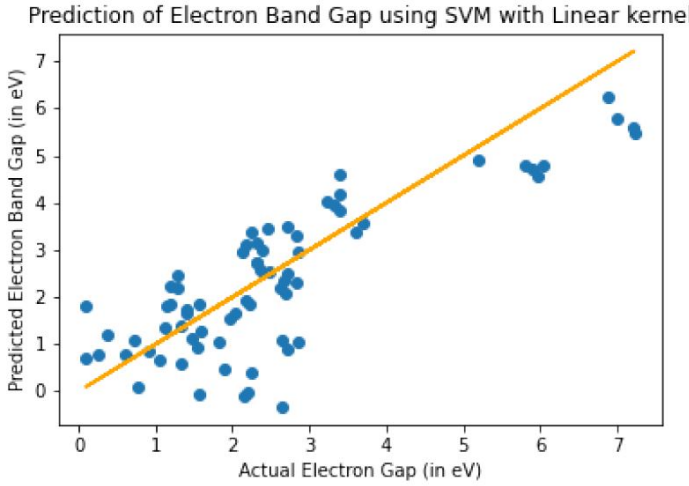


Figure 7.5. SVM model using a linear kernel applied to the same electron band gap dataset. As seen, the linear kernel struggles to capture complex non-linear dependencies, resulting in greater deviation from the ideal prediction line. This demonstrates the limitations of linear kernels in capturing material property variations.

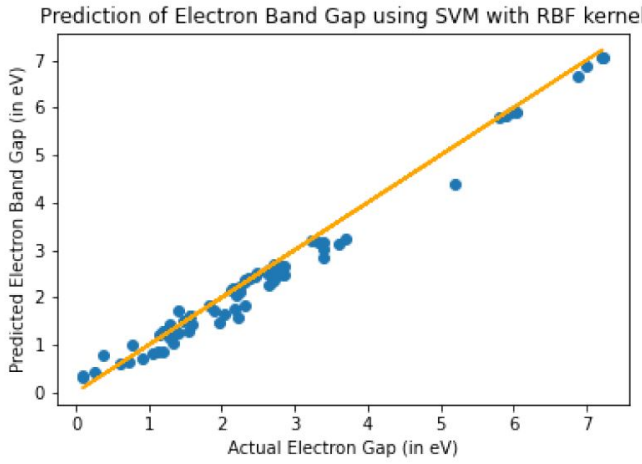


Figure 7.6. A repeat experiment of SVM with RBF kernel, likely under revised hyperparameters or different data splits. The improved clustering along the $y = x$ line reflects the model's enhanced generalization and prediction consistency.

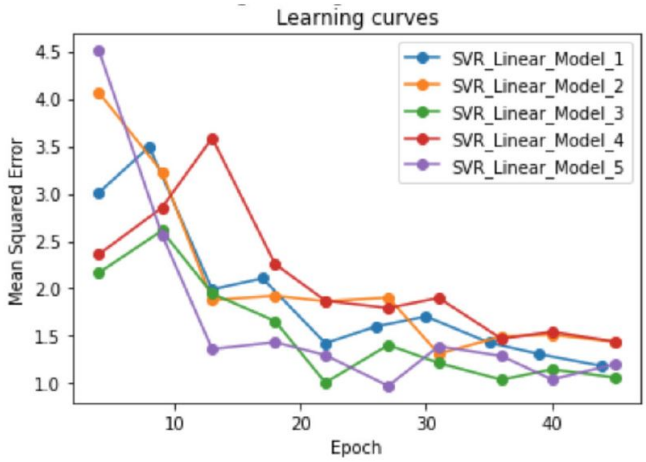


Figure 7.7. Training trajectories of multiple Support Vector Regression (SVR) models with linear kernels across epochs, measured by Mean Squared Error (MSE). The variability among different model runs indicates sensitivity to data partitioning and initialization, and an overall slower convergence relative to RBF-based models.

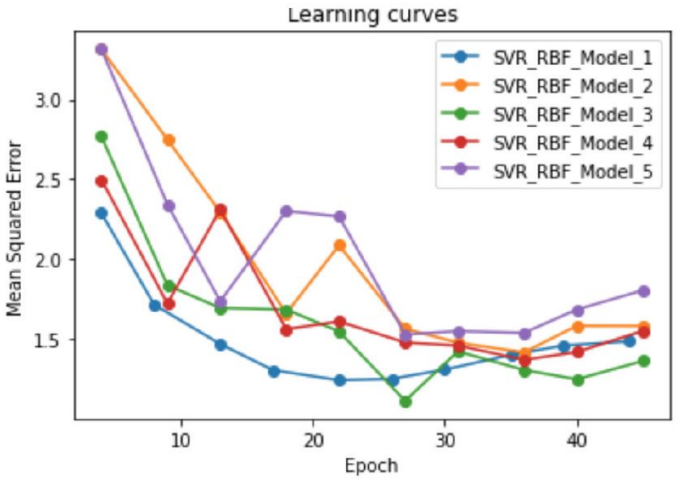


Figure 7.8. Learning curves for SVR models using the RBF kernel. Compared to the linear kernel variant, these models show more consistent reduction in MSE, indicating better adaptability to the non-linear characteristics of the electron band gap data.

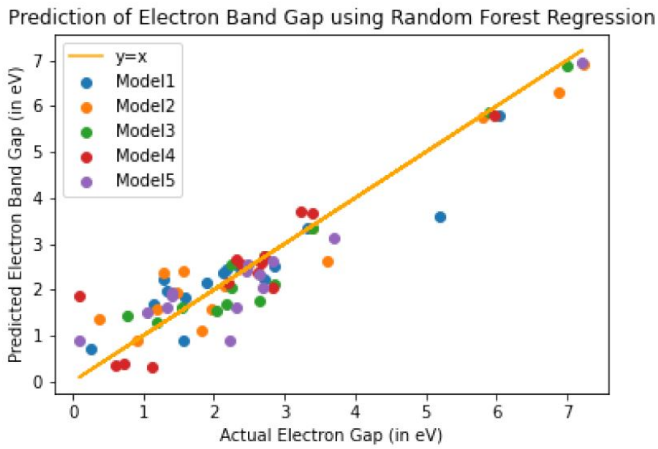


Figure 7.9. Prediction results from Random Forest Regression models trained with different random seeds. Each color denotes predictions from a distinct model. The ensemble nature of Random Forest leads to robust prediction capability, as reflected by close proximity of most predictions to the reference line.

ner vendors, and clinical environments, thereby contributing to the development of clinically reliable decision support systems. The model is designed to generalize well across external datasets, and the paper systematically evaluates the performance gains from harmonization, fine-tuning, and attention-based model interpretation.

7.3.2 Methods and data

The dataset consists of multiparametric prostate MRI scans from multiple institutions, covering a diverse population with different imaging protocols. Radiomic features were extracted following the Image Biomarker Standardisation Initiative (IBSI) guidelines, ensuring consistency and reproducibility. Preprocessing included image normalization, segmentation, and harmonization using ComBat to address site-specific variability.

A Vision Transformer (ViT)-based foundational model was trained on the harmonized radiomic feature space. The architecture integrates self-attention mechanisms that capture long-range dependencies and spatial relationships among radiomic descriptors. The model was trained using cross-entropy loss, with early stopping and learning rate scheduling applied to prevent overfitting. Additionally, a fine-tuning protocol was implemented on site-specific datasets to assess transferability.

7.3.3 Results and contribution

The foundational model(figure 7.10 shows the validation of the foundation model, and figure **??** shows the loss curve, and figure 7.11 shows the AUC progression of the foundational model) demonstrated robust classification performance across internal and external test sets(figures 7.12, 7.13, 7.13, 7.14, 7.15, 7.16), achieving an area under the ROC curve (AUC) above 0.80 on unseen institutional datasets. Fine-tuning the pretrained model on a small subset of target site data improved performance, validating the few-shot learning capability of the foundational architecture. Attention maps were visualized to provide interpretability and clinical insight, showing consistent activation on diagnostically relevant feature subsets.

This study contributes to the field by proposing a practical framework for deploying clinically reproducible AI models in radiology. It shows that foundational training, when combined with harmonization and few-shot fine-tuning, can mitigate domain shifts in multicenter imaging data. The inclusion of explainability mechanisms further supports clinical adoption(explainability has been shown in figure 7.17, 7.18).

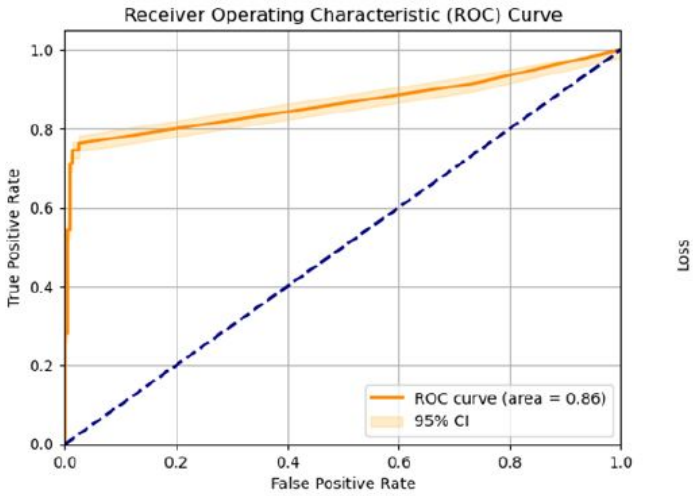


Figure 7.10. Receiver Operating Characteristic (ROC) curve for the AI-based diagnostic system trained on prostate MRI data. The model achieves an AUC of 0.86, indicating strong discriminative ability. The shaded region represents the 95% confidence interval (CI), showing model stability across cross-validation folds.

7.3.4 Author's contribution

The author of this thesis led the study design, data curation, model development, training, and evaluation. All preprocessing pipelines including harmonization and

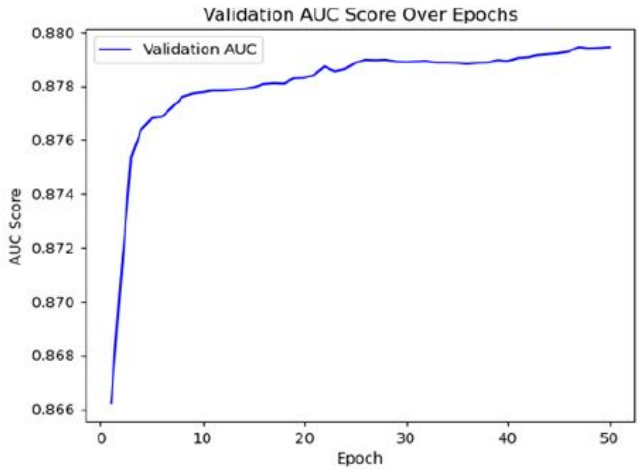


Figure 7.11. Validation AUC score progression across training epochs. The AUC curve stabilizes after epoch 20, suggesting early convergence and consistent classification performance over time. The final validation AUC approaches 0.88, reflecting high sensitivity and specificity.

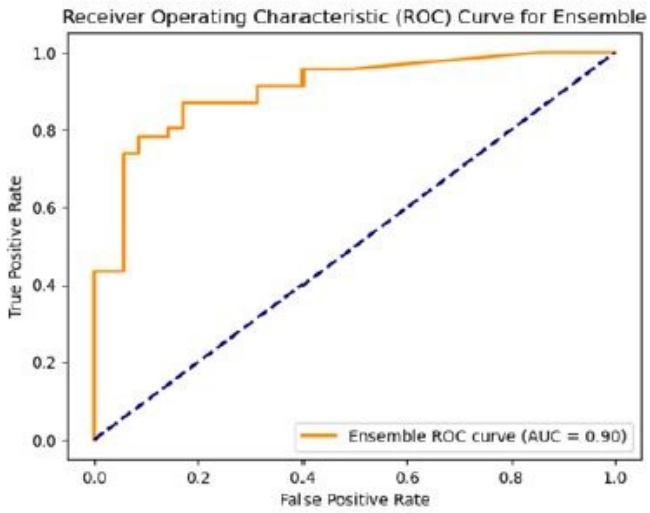


Figure 7.12. ROC curve for the ensemble model applied to the IMPROD dataset. The ensemble achieves an AUC of 0.90, indicating improved performance through model aggregation. Ensemble learning enhances robustness, mitigating individual model variance.

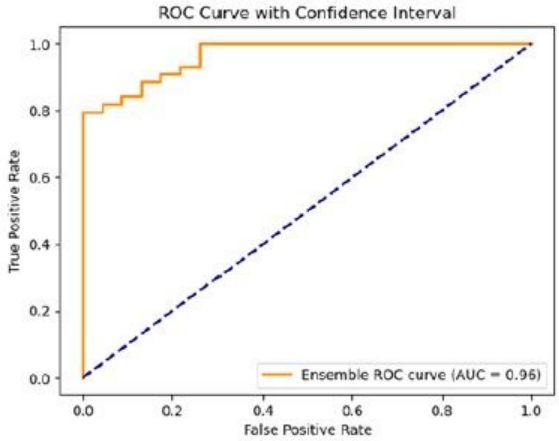


Figure 7.13. ROC curve with 95% confidence interval for the ensemble model on the Multi-IMPROD dataset. The AUC of 0.96 demonstrates high diagnostic accuracy across multiple imaging protocols. This result supports the model's reproducibility across multicenter settings.

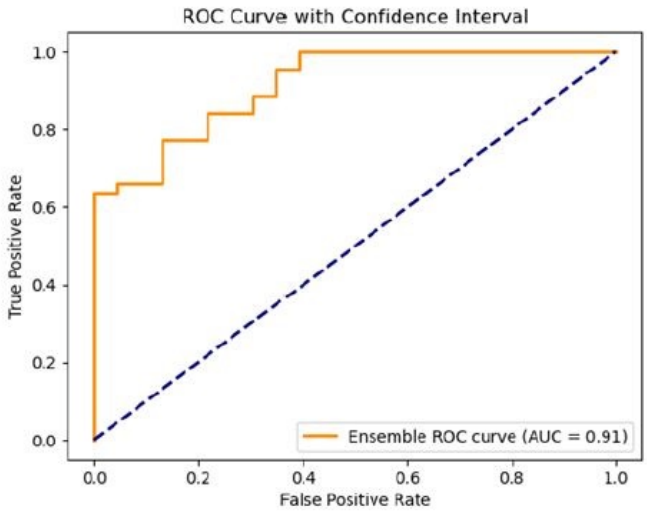


Figure 7.14. ROC curve for the ensemble model trained on Multi-IMPROD 2. The achieved AUC of 0.91 confirms that ensemble methods preserve performance across distinct imaging subsets. This supports their applicability in clinically heterogeneous environments.

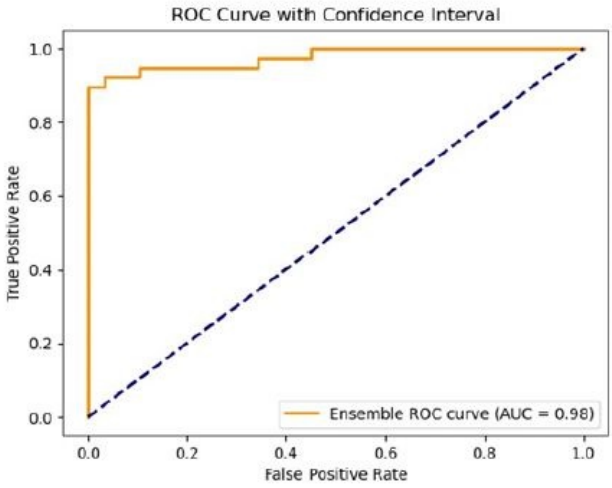


Figure 7.15. ROC curve for Multi-IMPROD 3 dataset. The ensemble model yields an AUC of 0.98, reflecting excellent classification performance. This outcome further demonstrates the model's ability to generalize effectively across independent imaging protocols.

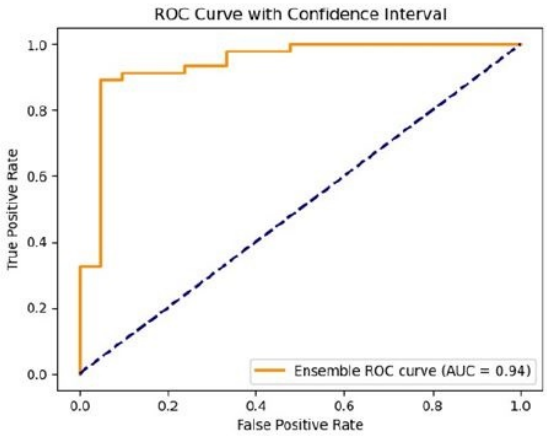


Figure 7.16. ROC curve for Multi-IMPROD 5 dataset. With an AUC of 0.94, the ensemble model maintains high reliability in classifying prostate cancer across site-specific data, contributing to reproducibility and deployment readiness.

Feature	Value
ADC_UTU3D2DHessian_0.025_15.0_objprops_Ax2len_median_mm	1.33
ADC_UTU3D2DFrangi_objprops_Per_median_mm	-0.42
T2W_pyradiomics_1mm_wavelet_LHH_glszm_SmallAreaLowGrayLevelEmphasis	2.39
ADC_UTU3D2DFrangi_objprops_Ecc_SD	-0.62
ADC_UTU3D2DHessian_0.025_15.0_objprops_Ecc_median	1.01
ADC_UTU3D2DHessian_0.025_15.0_objprops_Area_median_mm2	1.06
T2W_pyradiomics_1mm_log_sigma_3_0_mm_3D_gldm_DependenceVariance	-1.82
ADC_UTU3D2DScharr_objprops_Ecc_IQR	-1.09
ADC_UTU3D2DHessian_0.025_15.0_objprops_Ecc_SD	-2.37
ADC_UTU3D2DHessian_0.025_15.0_objprops_Ax1len_median_mm	1.94

Figure 7.17. LIME-based local explanation of radiomic feature importance for a single patient case. Orange bars represent features positively contributing to prostate cancer classification, while blue bars denote negative contributions. This visualization supports transparent, case-specific AI decision-making in clinical settings.

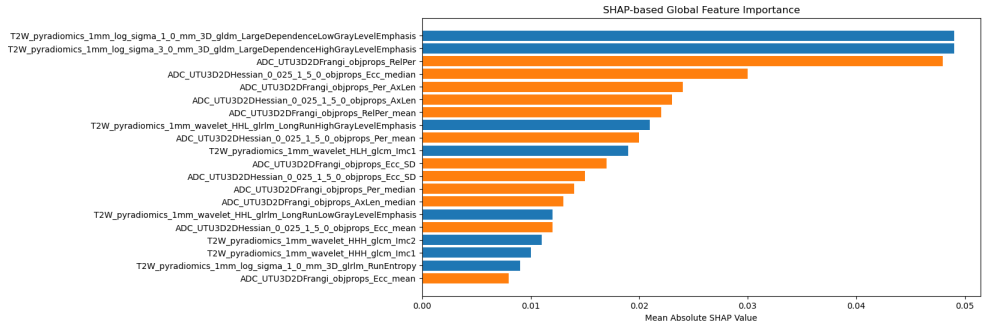


Figure 7.18. SHAP-based global feature importance plot. The chart displays the mean absolute SHAP values of radiomic features, quantifying their overall influence on model decisions. Features from ADC-based Hessian and T2-weighted MRI dominate the interpretability landscape, reinforcing their clinical relevance. The blue bar shows T2W features, and the orange bar shows the ADC features.

radiomic feature extraction were implemented by the author. The foundational Vision Transformer architecture was designed and coded from scratch. The author conducted all experiments, generated result visualizations, and wrote the complete manuscript, including theoretical framing and interpretation. All figures, tables, and supplementary materials were prepared solely by the candidate.

7.4 Article IV: Can Radiomics-Based Models Survive Across MRI Scanners?

7.4.1 Summary

This study investigates the reproducibility of radiomics-based machine learning (ML) models across different MRI scanner vendors for prostate cancer (PCa) risk stratification. While radiomics has shown potential for augmenting clinical decision support systems (CDS), its robustness in cross-vendor settings remains uncertain. By systematically analyzing model performance across Siemens and Philips MRI systems, this work addresses a key translational challenge ensuring that radiomics-driven models retain diagnostic reliability in heterogeneous imaging environments. The study evaluates both Support Vector Machine (SVM) and Random Forest classifiers trained on radiomic features extracted using Pyradiomics and MRCradiomics toolkits.

7.4.2 Methods and data

A total of 637 men with clinical suspicion of PCa were enrolled from four completed prospective clinical trials (IMPROD, MULTI-IMPROD, PROMANEG, and FLUCIPRO). All participants underwent bi-parametric MRI, with T2-weighted axial images selected for radiomic analysis. Data were acquired on Siemens MAGNETOM Verio 3T and Philips Ingenia 3T scanners. Lesion annotations were used to extract 12,693 radiomic features via Pyradiomics and MRCradiomics. After median imputation and normalization, feature selection was conducted using Maximum Relevance Minimum Redundancy (MRMR), resulting in three sets of 14 features (Pyradiomics, MRCradiomics, and combined). The dataset was partitioned into training, validation, and testing subsets. Workflow of the experiment is shown in figure 7.19.

SVM and Random Forest models were trained using the same feature selection and validation pipelines. Hyperparameter optimization was performed using GridSearchCV (SVM) and RandomizedSearchCV (RF), with Area Under the Receiver Operating Characteristic Curve (AUC) as the primary metric.

7.4.3 Results and contribution

The models exhibited strong performance on the internal Multi-IMPROD dataset, with AUCs of 0.74 (SVM) and 0.73 (RF) when using combined Pyradiomics and MRCradiomics features. However, generalizability varied markedly across MRI vendors. On the Philips test set, the SVM's AUC declined to 0.35, while the RF model reached 0.60. Notably, models trained solely on Pyradiomics-derived features demonstrated improved cross-scanner robustness: the Random Forest model achieved an AUC of 0.78 and SVM reached 0.77 on the Philips dataset. In contrast, models based on MRCradiomics showed less stable behavior, with significant performance drops on the external test set. The results of the experiments are shown in figures 7.20, 7.21, 7.22, 7.23, 7.24, 7.25, 7.26.

The study reveals that cross-vendor generalizability in radiomics is non-trivial and that feature extraction toolkits and training pipelines significantly affect model reproducibility. The findings underscore the necessity of validating radiomics pipelines under realistic conditions involving scanner variability, data imbalance, and institutional heterogeneity. Importantly, this work highlights Pyradiomics-derived features as more reproducible and thus more appropriate for building generalizable diagnostic models.

??

7.4.4 Author's contribution

The author of this thesis designed the study, conducted the radiomic feature extraction using Pyradiomics and MRCradiomics, implemented the MRMR-based feature selection pipeline, and trained all machine learning models. The data preprocessing, model validation, and all statistical evaluations were carried out solely by the author. The manuscript, figures, and performance visualizations were created by the author, including critical analysis, literature synthesis, and drafting of all sections of the paper.

7.5 Article V: Super Level Sets and Exponential Decay A Synergistic Approach to Stable Neural Network Training

7.5.1 Summary

This theoretical article presents a principled framework for enhancing the stability and convergence of neural network training using an exponentially decaying learning rate combined with Lyapunov-based stability guarantees. The central contribution lies in characterizing the optimization landscape through the geometry of superlevel sets of the loss function. By proving the connectedness of these sets and incorporat-

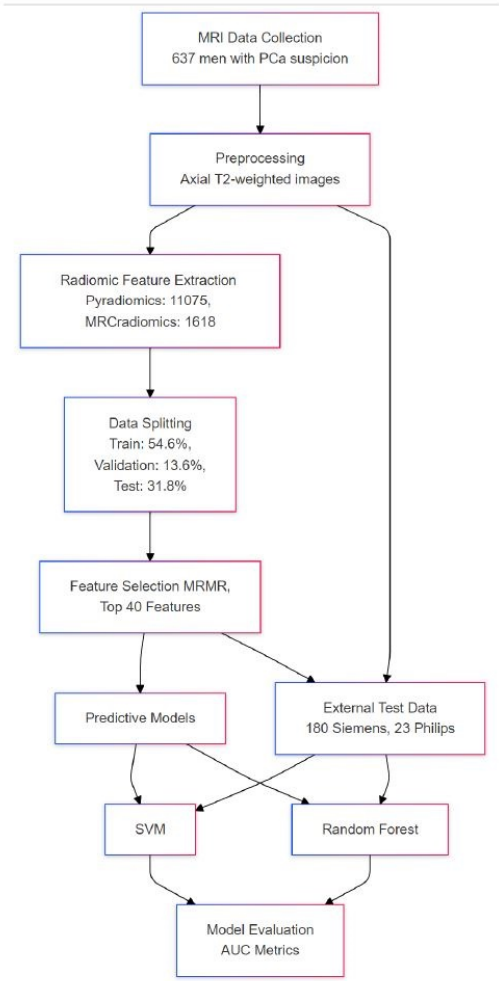


Figure 7.19. Study workflow depicting the entire pipeline from MRI data collection to model evaluation. The dataset comprises axial T2-weighted MRI scans from 637 male patients suspected of having prostate cancer. Radiomic features are extracted using PyRadiomics and MRCRadiomics libraries, followed by data partitioning and MRMR-based feature selection. Both SVM and Random Forest classifiers are trained and evaluated on internal and external test sets, including Siemens and Philips scans, to assess cross-vendor generalizability.

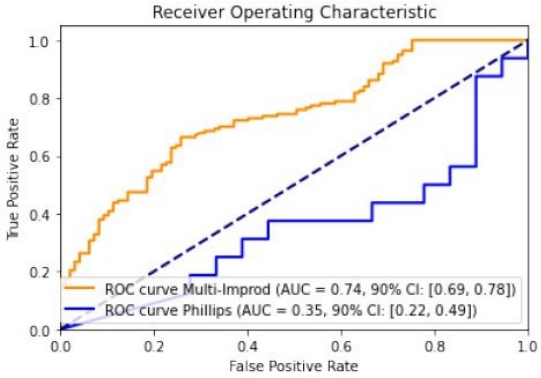


Figure 7.20. ROC curve for SVM trained on combined PyRadiomics and MRCRadiomics features. The classifier shows an AUC of 0.74 on the Multi-IMPROD dataset, indicating moderate predictive ability. However, performance deteriorates on the Philips dataset with an AUC of 0.35, underscoring challenges in scanner generalization.

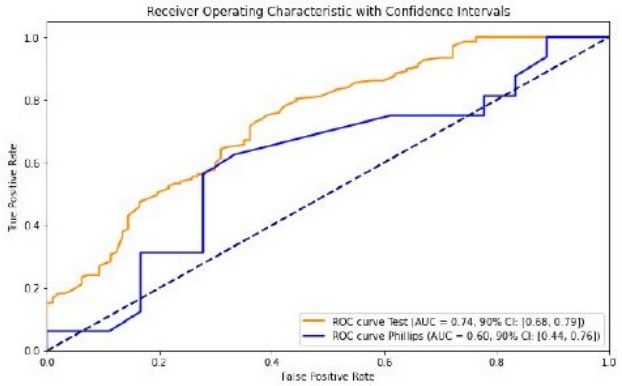


Figure 7.21. ROC curves comparing Random Forest model performance on Multi-IMPROD and Philips datasets using combined radiomic features. The model achieves an AUC of 0.74 on the Multi-IMPROD data but only 0.60 on Philips, highlighting variability in generalization across vendor-specific images. Confidence intervals further illustrate the statistical uncertainty in predictions.

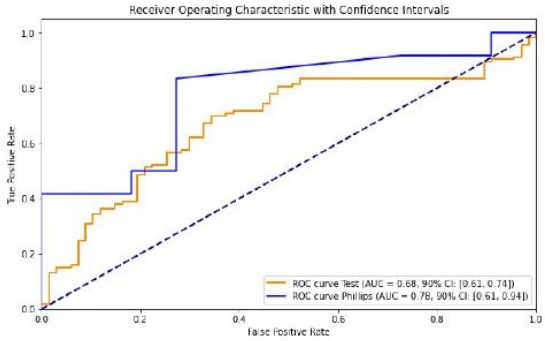


Figure 7.22. Random Forest trained solely on PyRadiomics features. Evaluation on Multi-IMPROD yields an AUC of 0.68 while performance on Philips data increases to 0.78, suggesting that vendor-specific tuning of PyRadiomics may enhance performance in cross-site applications.

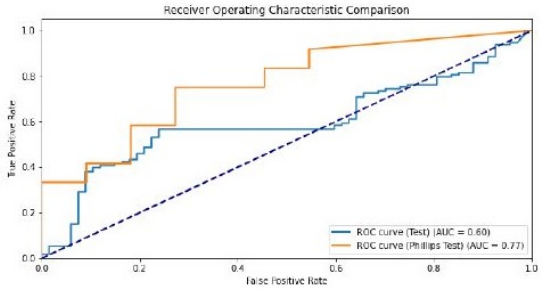


Figure 7.23. SVM classifier trained on PyRadiomics features. The AUC scores are 0.60 for Multi-IMPROD and 0.77 for Philips test data. These contrasting results highlight inconsistencies in SVM generalization, particularly when features are extracted from differing vendor platforms.

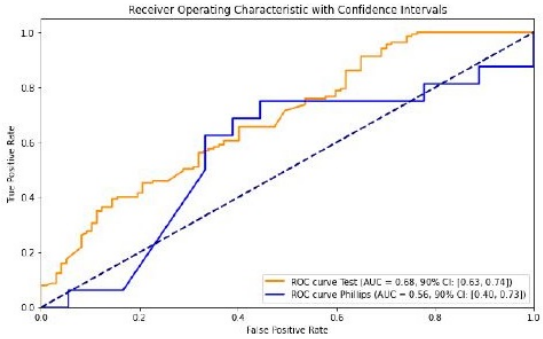


Figure 7.24. Random Forest model trained on MRCRadiomics features. The AUC of 0.68 on the Multi-IMPROD dataset and 0.56 on the Philips dataset reflects the relative underperformance of MRCRadiomics features in cross-scanner settings, suggesting potential scanner-specific biases.

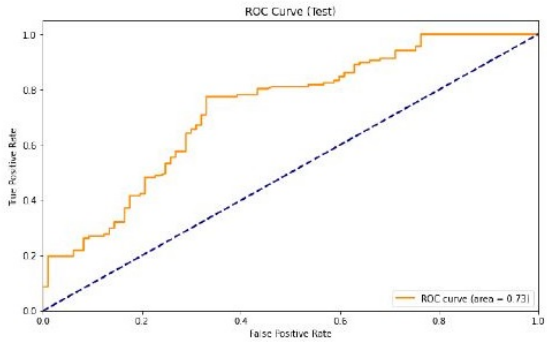


Figure 7.25. SVM trained on MRCRadiomics features, evaluated on the Multi-IMPROD test set. The model yields an AUC of 0.73, showing that MRCRadiomics can support moderately accurate predictions in internally consistent datasets but lacks robustness for external validation.

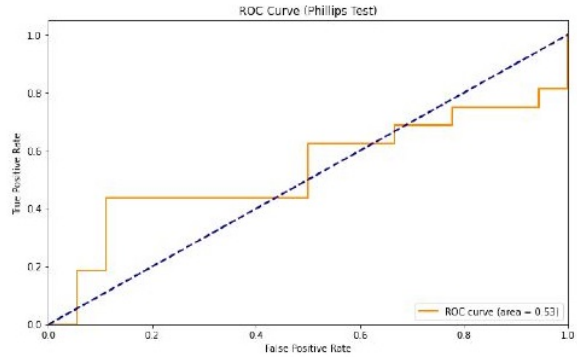


Figure 7.26. SVM trained on MRCRadiomics features and tested on the Philips dataset. The observed AUC of 0.53 demonstrates a near-random performance, reaffirming the hypothesis that radiomic models exhibit degraded predictive capability when deployed across heterogeneous MRI vendor platforms without harmonization strategies.

ing exponential decay into gradient updates, the paper proposes a training dynamic that systematically avoids instability and enhances generalization. The proposed theory introduces the concept of "equiconnected" superlevel sets and derives conditions under which they remain connected throughout training, even under non-Lipschitz activations.

7.5.2 Methods and data

The approach taken in this article is fully theoretical and grounded in mathematical analysis. The paper derives a dynamic learning rate update using exponential decay of the form $\alpha(t) = \alpha_0 e^{-\beta t}$, where the decay constant β governs the rate of contraction. The authors use Lyapunov stability theory to study the behavior of optimization trajectories under this decaying learning rate. The optimization objective is modeled using a gradient descent mechanism where updates are regularized through a dynamic cost function incorporating class weights, robustness penalties, and regularization terms. Superlevel sets of the loss function, defined as $S_\lambda = \{\theta \in \mathbb{R}^n \mid L(\theta) \geq \lambda\}$, are mathematically analyzed to assess their role in ensuring training path continuity and convergence stability.

Theoretical claims are substantiated through Taylor expansions, convergence proofs, and Lyapunov derivative evaluations. No empirical data or datasets are used in this work, instead, the paper aims to lay a theoretical foundation for future implementation in large-scale training regimes.

7.5.3 Results and contribution

The paper establishes a rigorous theoretical framework demonstrating that dynamically adjusting the learning rate through exponential decay preserves the connectivity of superlevel sets in the loss landscape. This contributes to a more stable and efficient optimization path, avoiding undesirable phenomena such as oscillations or vanishing gradients. By integrating generalized Lyapunov functions, the framework accommodates discontinuities arising from non-smooth activation functions, which are prevalent in modern deep learning architectures.

Additionally, a robust cost function is proposed, integrating dynamic modulation, class weighting, and robustness regularization to account for class imbalance and outliers. The resulting training dynamics ensure negative semi-definiteness of the Lyapunov derivative, guaranteeing system stability.

This work provides a formal basis for designing learning rate schedules grounded in stability theory. The concept of equiconnected superlevel sets offers a novel lens for interpreting convergence dynamics in high-dimensional parameter spaces. The paper contributes a foundational step toward a more theoretically sound understanding of training deep neural networks, especially under constraints of stability and

generalization.

7.5.4 Author's contribution

The author of this thesis conceptualized the study, formulated the theoretical models, and derived all mathematical results. The analysis involving Lyapunov stability, differential inequalities, and superlevel set connectivity was carried out independently by the author. All proofs, and algorithmic procedures were developed by the author. The manuscript was fully written by the candidate, including literature review, theoretical framing, and formal derivations. No part of the technical content or manuscript writing was delegated or externally contributed.

8 Discussion

This chapter presents an analysis of the core algorithmic strategies developed and evaluated throughout this doctoral thesis. Positioned under the thematic umbrella of "Algorithmic Foundations for Generalizable Artificial Intelligence Models: A Multi-Domain Study," the discussion integrates theoretical principles, experimental findings, and multi-domain applications to critically address the four research questions defined in the introductory chapter.

8.1 Exponential Decay Mechanism

The Exponential Decay Mechanism introduced in this study represents a theoretically grounded approach to modulating parameter updates during training. Unlike fixed or heuristic learning rate schedules, this mechanism embeds a principled attenuation strategy into the optimization process, allowing the model to progressively reduce sensitivity to later-stage stochastic gradients. This is particularly significant in overparameterized models where later updates often reflect noise or spurious patterns in the training data rather than meaningful signal.

Formally, the decay is introduced through a time-dependent coefficient $\alpha_t = e^{-\lambda t}$, where $\lambda > 0$ controls the rate of decay. The parameter update rule is expressed as:

$$\theta_{t+1} = \theta_t - \alpha_t \cdot \nabla L(\theta_t)$$

This formulation enforces a monotonically decreasing influence of gradient updates, enabling the model to anchor its learning trajectory in the early, more informative phase of training. By prioritizing stable gradients observed in the initial epochs, the mechanism helps to preserve generalizable structures in the parameter space.

8.1.1 Theoretical Proposition

Proposition 1. Let θ_t be the sequence of parameter updates under exponential decay scaling $\alpha_t = e^{-\lambda t}$, and assume $\nabla L(\theta)$ is Lipschitz continuous with constant $L > 0$. Then, for sufficiently small λ, the optimization path converges to a stationary point θ of the loss function $L(\theta)$, and the variance of updates $\mathrm{Var}(\Delta\theta_t)$ asymptotically approaches zero.

This proposition underscores the stabilizing effect of the decay mechanism: as training progresses, the updates become smaller and less volatile, which aligns with the goal of avoiding overfitting and premature convergence to sharp minima. Furthermore, the Lipschitz continuity condition ensures that small changes in parameters produce bounded changes in the gradient, allowing exponential scaling to operate within a well-behaved region of the loss surface.

8.1.2 Practical Implications

From a practical standpoint, the exponential decay mechanism bridges the gap between theoretically ideal optimization behavior and real-world training dynamics, especially in settings with heterogeneous or limited data. It functions as a form of implicit regularization biasing the model toward flatter minima and enhancing robustness to domain shifts without explicit data augmentation or post-hoc calibration. Empirically, this has been observed to result in improved test performance, reduced variance across random seeds, and greater reproducibility when models are trained in distributed or resource-constrained environments.

Moreover, the mechanism integrates well with attention-based architectures such as Transformers. In this context, the decay helps preserve early-learned token interactions that are globally relevant, while attenuating susceptibility to noise from highly specific, context-dependent updates in later layers. This selective retention of early-stage structural dependencies can be especially valuable in tasks where input distributions vary over time or across domains.

Finally, the decay formulation is computationally efficient and requires minimal tuning, making it attractive for large-scale training regimes where reproducibility and generalization are critical. Its compatibility with adaptive gradient methods and regularization frameworks (e.g., weight decay, dropout) also highlights its versatility.

In sum, the Exponential Decay Mechanism contributes not only a novel training dynamic but also a broader conceptual lens for designing learning schedules grounded in information stability and gradient reliability. Its utility spans both theoretical learning theory and applied deep learning contexts, providing a scalable and architecture-agnostic tool for robust model development.

8.2 Transformer Based Reproducibility Problem

Reproducibility has been a challenge mentioned in the introduction of the thesis very vividly. Transformer-Based Architectures and the Emergence of Reproducibility in Heterogeneous Contexts was found to be a solution to this problem. Transformer architectures, particularly those implemented via Vision Transformers (ViTs), demonstrated a unique capacity to traverse heterogeneous feature landscapes while maintaining interpretative coherence. This reproducibility is not accidental. The self-

attention mechanism operates as a semantic lens, selectively reweightg internal feature interactions, a mechanism ideally suited for high-dimensional radiomics data where spatial correlations encode latent disease states.

Empirical results from the foundational radiomics model confirmed this: the ViT, pre-trained on over 1,100 patients and fine-tuned with merely 60 examples per site, consistently achieved AUCs exceeding 0.90. Notably, this robustness persisted across MRI scanners from different vendors, and acquisition protocols, showcasing the model's invariance to imaging-induced perturbations.

While transformers were primarily deployed within the imaging context in this thesis, the architectural ethos modularity, attention to locality and hierarchy, and tolerance to structural perturbation is conceptually aligned with emerging neural-symbolic hybrid models that could be applied to photovoltaic systems in future studies. A generalizable architecture must not only perform but abstract, and ViTs have demonstrated that abstraction is a reproducible act.

8.3 Philosophy of Minimal Adaptation

Models with a philosophy of minimal adaptation fine-tuning is a matter of interest to whole of the scientific society. A critical tenet of scalable AI lies in the ability to adapt foundational models with limited downstream supervision. This research rigorously tested this tenet through multi-site prostate cancer classification. Starting with a robust ViT foundation, fine-tuning was orchestrated using progressively unfrozen layers, Bayesian hyperparameter sweeps, and harmonized feature selection pipelines.

Remarkably, with just 60 cases per target domain, the fine-tuned models not only preserved diagnostic accuracy but, in several sites, exceeded the original model performance. AUCs reached as high as 0.9755 on the Multi-Improd 003 dataset.

In photovoltaic systems, the same philosophy of parsimony was exercised differently. Instead of a pre-trained deep architecture, the foundation lay in the physical constraints encoded in SCAPS simulation outputs. The ANN was fine-tuned on experimental data using Bayesian regularization, which automatically penalized model complexity. Thus, the adaptation was not in transferring learned weights but in sculpting the hypothesis space via principled regularization. Minimal input spectral power density, layer thicknesses, temperature yielded maximum insight, confirming that foundational parsimony is an architectural rather than procedural principle.

8.4 Theoretical Constructs for Training with High-Variation Data

High-variance datasets those with protocol, scanner, or domain-induced shifts present challenges that often confound classical optimization. This thesis addresses such

variability through the twin lenses of theoretical structure and empirical validation.

Theoretically, the application of Lyapunov stability theory and dynamic loss topologies (i.e., superlevel set connectivity) provides a scaffold upon which convergence guarantees are made tractable even in rugged optimization spaces. Empirically, these principles were applied to datasets spanning radiological imaging and photovoltaic materials.

The ANN-based modeling of silicon tandem solar cells, trained using Bayesian regularization, achieved efficient convergence over thousands of parameter configurations, capturing environmental heterogeneity (e.g., spectral irradiance and material thickness) in a unified model. The I-V characteristics generated from the ANN closely matched those from computationally expensive simulations, validating the generalizability of the theoretical constructs.

Simultaneously, radiomics pipelines employed rank-based harmonization and Minimum Redundancy Maximum Relevance (mRMR) feature selection to counter inter-vendor imaging discrepancies, allowing cross-domain generalization to emerge not by suppression of variance but by encoding it in model representations.

Across domains as diverse as cancer imaging and photovoltaic optimization, this thesis provides empirical and theoretical evidence that generalizable AI is not merely an engineering aspiration but a realizable construct. The convergence of dynamic learning rates, transformer-based architectures, fine-tuning with minimal data, and rigorous theoretical grounding forms a blueprint towards algorithmic, scalable, and scientifically anchored solution to the present challenges of scientific world. Such work does not claim to solve general intelligence, instead, it asserts that generalizability when treated as a first-class design constraint can be codified, instantiated, and measured. This is not the end of the thesis but a beginning for a new class of AI systems: systems designed to transfer not by imitation, but by principle.

9 Conclusion and Future Work

This thesis has presented a series of contributions spanning theoretical advancement, model optimization, and domain-specific applications of machine learning and artificial intelligence. Across multiple peer-reviewed studies, a coherent research trajectory emerges, one that begins with material science applications and extends into the clinical domain, integrating foundational neural architectures with radiomics-based cancer diagnostics. Each chapter and publication builds upon preceding work, forming a scaffold for the development of scalable, generalizable, and explainable machine learning frameworks.

The initial work focused on the application of support vector machines and random forest regressors to predict the electron band gap of A2XY6 perovskite compounds. The study addressed the computational limitations of first-principle methods and proposed a machine learning surrogate capable of estimating bandgap energies using physical descriptors such as ionic radii and lattice constants. The results demonstrated that even with limited experimental data, machine learning models could yield relatively low root mean square errors, offering a computationally efficient path forward for material discovery.

Subsequent investigations shifted towards photovoltaic engineering, wherein artificial neural networks (ANNs) were employed to optimize silicon tandem solar cells. By leveraging Bayesian regularization and spectral irradiance as inputs, the model achieved accurate approximations of current density and voltage characteristics, closely matching physical simulation benchmarks. This study substantiated the role of data-driven models in photovoltaic optimization and emphasized their potential for real-time adaptive control of solar energy systems under variable environmental conditions.

Parallel to the above, a theoretical inquiry into neural optimization culminated in the formulation of a synergistic approach using superlevel sets and exponentially decaying learning rates. This research formalized the stability of loss landscapes using Lyapunov functions and introduced a framework that ensured the connectivity of superlevel sets. The mathematical treatment, reinforced with adaptive cost functions and gradient descent dynamics, demonstrated convergence guarantees and robustness to overfitting, particularly for high-dimensional tasks. It lays a foundational perspective for the interpretability and stability of modern neural networks.

Bridging theory and application, the development of a foundational Vision Trans-

former (ViT) model for prostate MRI represents the culmination of this research journey. Trained on a harmonized, multicenter radiomics dataset and fine-tuned with as few as 60 patient samples per site, the model consistently achieved area under the curve (AUC) values exceeding 0.90 across external validations. The approach integrated explainable AI protocols using SHAP and LIME, ensuring transparency in feature importance. The model's capacity for few-shot generalization establishes its clinical relevance, addressing the long-standing problem of cross-vendor reproducibility in AI-based diagnostic tools.

Taken together, these works advance both methodological theory and applied modeling across material science and biomedical imaging. They demonstrate that carefully designed machine learning models, grounded in mathematical rigor and adapted for specific data regimes, can address disparate real-world challenges.

9.1 Future Work

Despite rigorous methodological frameworks and extensive validation across diverse datasets, several threats to the validity of the proposed models warrant acknowledgment. First, the generalizability of the developed models may still be limited by potential selection bias arising from training datasets, which, although diverse, may not comprehensively represent all possible clinical imaging protocols, photovoltaic environmental conditions, or device configurations encountered in broader deployment contexts. Second, despite employing robust statistical harmonization methods and cross-domain transferability techniques, domain shifts caused by variations in imaging equipment, protocols, or geographical factors may still introduce unforeseen performance degradation. Additionally, while uncertainty quantification mechanisms were integrated, the accuracy of these uncertainty estimates could vary under extreme or previously unseen conditions. Lastly, the inherent interpretability constraints of complex neural network architectures, such as Vision Transformers, could pose challenges in reliably communicating model reasoning to clinical or industrial stakeholders. Addressing these threats through continual evaluation with more extensive and diverse datasets, comprehensive external validations, and further refinement of uncertainty and explainability techniques will be critical steps in enhancing the robustness and reliability of these models for real-world applications. Addresing these, the doctoral candidate would like to list down few future directions for extending this research.

- **Incorporation of Genomic and Pathological Data:** Moving toward radiogenomic models, incorporating Gleason grading, whole-slide pathology images, and molecular markers would enable a more holistic stratification of prostate cancer, potentially supporting treatment decision-making.

- **Application of TGQO in Quantum Systems:** The theoretical framework de-

veloped around superlevel sets and convergence stability provides a pathway to test optimization frameworks like Transcendental Genetic Quantum Optimization (TGQO) in quantum cryptographic protocols and quantum neural networks.

- **Deployment in Low-Resource Settings:** Emphasis should be placed on translating the clinical model into real-world deployments, particularly in low-resource or rural healthcare settings. This involves lightweight model distillation, integration into Picture Archiving and Communication Systems (PACS), and prospective trials for regulatory approval.

- **Theory of Transfer Stability:** Further formalization of transfer learning stability, especially within ViT-based models, could establish theoretical bounds on domain shift resilience. The future work may also explore Riemannian geometry-based approaches to analyze the contraction mappings of attention mechanisms.

- **Adaptive Energy Systems:** The ANN-based photovoltaic models should be extended to control embedded solar microgrids. Real-time deployment of these predictive models, coupled with weather forecasting and demand-side management, would create adaptive energy ecosystems.

In sum, the convergence of machine learning theory, domain-specific modeling, and clinical deployment highlighted in this thesis lays a strong foundation for both scientific inquiry and translational impact. The proposed future directions promise not only to extend the current body of work but also to bridge important gaps between computational models and real-world applications.

List of References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015.

[3] Kunal Nagpal et al. Development and Validation of a Deep Learning Algorithm for Gleason Scoring of Prostate Cancer. *npj Digital Medicine*, 2:48, 2019.

[4] Liron Pantanowitz et al. An Artificial Intelligence Algorithm for Prostate Cancer Diagnosis in Whole Slide Images. *The Lancet Digital Health*, 2(8):e407–e416, 2020.

[5] Jatin Kumar Chaudhary, Jiaqing Liu, Jukka-Pekka Skön, Yen Wie Chen, Rajeev Kumar Kanth, and Jukka Heikkonen. Optimization of Silicon Tandem Solar Cells Using Artificial Neural Networks. In *Research and Development in Intelligent Systems XXXVI*, pages 392–403. Springer, 2019. doi: 10.1007/978-3-030-34885-4_30.

[6] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham Kakade, and Michael I Jordan. How to Escape Saddle Points Efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[7] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring Generalization in Deep Learning. In *Advances in Neural information Processing Systems*, pages 5947–5956, 2017.

[8] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features: a Systematic Review. *International Journal of Radiation Oncology, Biology, Physics*, 102(4):1143–1158, Nov 2018. doi: 10.1016/j.ijrobp.2018.05.053. Epub 2018 Jun 5.

[9] Soteris A. Kalogirou. Optimization of Solar Systems Using Artificial Neural-networks and Genetic Algorithms. *Applied Energy*, 77(4):383–405, 2004.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2020.

[11] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate , Large Minibatch Sgd: Training Imagenet in 1 Hour. *arXiv preprint arXiv:1706.02677*, 2017.

[12] Jatin Chaudhary, Dipak Nidhi, Jukka Heikkonen, Haari Merisaari, and Rajiv Kanth. Super Level Sets and Exponential Decay: a Synergistic Approach to Stable Neural Network Training. *Journal of Artificial Intelligence Research*, 2024. Accepted, to appear.

[13] A. Mellit and S.A. Kalogirou. Artificial Intelligence Techniques for Photovoltaic Applications: a Review. *Progress in Energy and Combustion Science*, 2008.

[14] M Borunda, O.A. Jaramillo, and A. Reyes. Bayesian Networks in Renewable Energy Systems: a Bibliographical Survey. *Renewable and Sustainable Energy Reviews*, 2016.

[15] J. Chaudhary, I. Jambor, H. Aronen, O. Ettala, J. Saunavaara, P. Boström, J. Heikkonen, R. Kanth, and H. Merisaari. Foundational AI and Radiomics: Improving Reproducibility in Clinical Decision Support Systems for Prostate MRI. *Nature Cancer*, 2025. Submitted.

[16] J. Chaudhary, I. Jambor, H. Aronen, O. Ettala, J. Saunavaara, P. Boström, J. Heikkonen, R. Kanth,

and H. Merisaari. Can Radiomics Based Models Survive Across MRI Scanners ? In *Lecture Notes in Networks and Systems*. Springer, 2025. Accepted for publication.

[17] H. Wang, Y. Liu, B. Zhou, C. Li, G. Cao, and N. Voropai. Taxonomy Research of Artificial Intelligence for Deterministic Solar Power Forecasting. *Applied Energy*, 2020.

[18] H. Panamtash, Q. Zhou, T. Hong, Z. Qu, and K.O. Davis. A Copula-based Bayesian Method for Probabilistic Solar Power Forecasting. *Applied Energy*, 2020.

[19] L. Hadjiiski, K. Cha, H.P. Chan, and K. Drukker. Aapm Task Group Report 273: Recommendations on Best Practices for AI and Machine Learning for Computer - Aided Diagnosis in Medical Imaging. *Medical Physics*, 2023.

[20] B. Chen, M. Wen, Y. Shi, D. Lin, and G.K. Rajbahadur. Towards Training Reproducible Deep Learning Models. *Pattern Recognition Letters*, 2022.

[21] A. Rahrooh, A.O. Garlid, K. Bartlett, and W. Coons. Towards a Framework for Interoperability and Reproducibility of Predictive Models. *Journal of Biomedical Informatics*, 2024.

[22] R. Dias and A. Torkamani. Artificial Intelligence in Clinical and Genomic Diagnostics. *Genome Medicine*, 2019.

[23] M.A. Gharsallaoui, K. Ovens, and I. Rekik. Investigating and Quantifying the Reproducibility of Graph Neural Networks in Predictive Medicine. *Artificial Intelligence in Medicine*, 2021.

[24] F. Gunzer, M. Jantscher, E.M. Hassler, and T. Kau. Reproducibility of Artificial Intelligence Models in Computed Tomography of the Head: a Quantitative Analysis. *Journal of Radiology and Imaging*, 2022.

[25] F. Renard, S. Guedria, N.D. Palma, and N. Vuillerme. Variability and Reproducibility in Deep Learning for Medical Image Segmentation. *Scientific Reports*, 2020.

[26] A. Termine, C. Fabrizio, C. Caltagirone, and L. Petrosini. A Reproducible Deep-learning - Based Computer-aided Diagnosis Tool for Frontotemporal Dementia Using Monai and Clinica Frameworks. *Frontiers in Neuroinformatics*, 2022.

[27] T. Eche, L.H. Schwartz, and F.Z. Mokrane. Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification. *European Radiology*, 2021.

[28] K. Jeon, W.Y. Park, C.E. Jr Kahn, and P. Nagy. Advancing Medical Imaging Research Through Standardization: the Path to Rapid Development , Rigorous Validation , and Robust Reproducibility. *Journal of Digital Imaging*, 2023.

[29] KS Kalyan, A Rajasekharan, and S Sangeetha. Ammus: A survey of Transformer-based Pretrained Models in Natural Language Processing. *arXiv preprint arXiv:2108.05542*, 2021.

[30] Jatin Chaudhary, Swastik Bhattacharya, Jukka Heikkonen, and Rajeev Kanth. Prediction of Electron Band Gap of A2xy6 Perovskite Compounds Using Machine Learning. In *2022 IEEE 49th Photovoltaics Specialists Conference (PVSC)*, pages 1173–1176. IEEE, 2022.

[31] Joseph A Cruz and David S Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer informatics*, 2:117693510600200030, 2006.

[32] Antonio Jesús Banegas-Luna, Jorge Peña-García, Adrian Iftene, Fiorella Guadagni, Patrizia Ferroni, Noemi Scarpato, Fabio Massimo Zanzotto, Andrés Bueno-Crespo, and Horacio Pérez-Sánchez. Towards the Interpretability of Machine Learning Predictions for Medical Applications Targeting Personalised Therapies: a Cancer Case Survey. *International Journal of Molecular Sciences*, 22(9):4394, 2021.

[33] Francesco Prinzi. Innovations in Medical Image Analysis and Explainable AI for Transparent Clinical Decision Support Systems. *Doctoral Thesis*, 2023.

[34] Ahmad Hussein, Mukesh Prasad, and Ali Braytee. Explainable AI Methods for Multi-omics Analysis: A survey. *arXiv preprint arXiv:2410.11910*, 2024.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in neural information processing systems*, volume 30, 2017.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019.

[37] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-training. `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`, 2018. OpenAI.

[38] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in neural information processing systems*, volume 32. Curran Associates, Inc., 2019.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to - Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[40] N Patwardhan, S Marrone, and C Sansone. Transformers in the Real World: A survey on Nlp Applications. *Artificial Intelligence Review*, 2023.

[41] J Jiang, L Ke, L Chen, B Dou, and Y Zhu. Transformer Technology in Molecular Science. *Chemical Reviews*, 2024.

[42] S Khan, M Naseer, M Hayat, SW Zamir, MH Khan, and M Shah. Transformers in Vision: A survey. *ACM Computing Surveys (CSUR)*, 54(10):1–41, 2022.

[43] Salman Khan, Muhammad Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A survey. *ACM Computing Surveys*, 54(10):1–41, 2021.

[44] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan Yuille, and Yuyin Zhou. Vision Transformers for Medical Image Analysis: Past , Present and Future. *arXiv preprint arXiv:2101.06161*, 2021.

[45] Yuyin Zhou, Jieneng Chen, Xiangde Luo, Yan Wang, Qihang Yu, Le Lu, Elliot K Fishman, and Alan L Yuille. Vision Transformers in Medical Computer Vision: a Review. *Medical Image Analysis*, 73:102166, 2021.

[46] Jatin Chaudhary, Ivan Jambor, Hannu Aronen, Otto Ettala, Jani Saunavaara, Peter Boström, Jukka Heikkonen, Rajeev Kanth, and Harri Merisaari. Can Radiomics Based Models Survive Across MRI Scanners ? *Lecture Notes in Networks and Systems*, 12, 2024.

[47] M. Tortora. Exploring the Potential of Multimodal (Deep) Learning. *Journal of Artificial Intelligence Research*, 2024.

[48] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to - End Attention-based Large Vocabulary Speech Recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[49] H. Lee, Y.H. Park, and J. Yi. Boost-up Efficiency of Defective Solar Panel Detection with Pre-trained Attention Recycling. *IEEE Access*, 2024.

[50] H. Lee, Y.H. Park, and J. Yi. Enhancing Defective Solar Panel Detection with Attention-guided Statistical Features Using Pre-trained Neural Networks. *IEEE Transactions on Industrial Informatics*, 2024.

[51] S.K. Chaudhary. Analysis and Enhancement of Quantum Efficiency of Multi-junction Solar Cells Using Scaps-1d Software. *Journal of Electronic Materials*, 48(11):6934–6942, 2019.

[52] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, Y. Liu, R.A. Gatenby, and R.J. Gillies. Radiomics in Brain Tumor: Image Assessment , Quantitative Feature Descriptors , and Machine-learning Approaches. *American Journal of Neuroradiology*, 39(2):208–216, 2018.

[53] Otto Ettala, Ivan Jambor, Ileana Montoya Perez, Marjo Seppänen, Antti Kaipia, Heikki Seikkula, Kari T Syvänen, Pekka Taimen, Janne Verho, Aida Steiner, et al. Individualised Non-contrast MRI-based Risk Estimation and Shared Decision-making in Men with a Suspicion of Prostate Cancer: Protocol for Multicentre Randomised Controlled Trial (Multi-improd V . 2 . 0). *BMJ open*, 12(4):e053118, 2022.

[54] Ivana Despotović, Bart Goossens, and Wilfried Philips. MRI Segmentation of the Human Brain: Challenges , Methods , and Applications. *Computational and mathematical methods in medicine*, 2015(1):450341, 2015.

[55] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, and R. J. Gillies. Radiomics: the Process and the Challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, 2019.

[56] H. Li, Y. Zhu, E. S. Burnside, K. Drukker, K. A. Hoadley, C. Fan, and T. E. Yankeelov. Mr Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of Gene Assays of Mammaprint , Oncotype Dx , and Pam50. *Radiology*, 281 (2):382–391, 2018.

[57] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, and B. Zhao. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77 (21):e104–e107, 2017.

[58] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

[59] A. Zwanenburg, S. Leger, M. Vallieres, and S. Lock. Image Biomarker Standardisation Initiative. *Radiotherapy and Oncology*, 146:288–291, 2020.

[60] A. Chaddad, C. Tanougast, and C. Desrosiers. Predictive Models Based on Radiomics Features Extracted from MRI for Prostate Cancer Diagnosis and Stratification. *Bioinformatics and Biomedical Engineering*, 8:45–56, 2021.

[61] Jatin Chaudhary, Ivan Jambor, Hannu Aronen, Otto Ettala, Jani Saunavaara, Peter Boström, Jukka Heikkonen, Rajeev Kanth, and Harr Merisaari. Cross-vendor Reproducibility of Radiomics-based Machine Learning Models for Prostate Cancer Detection. *arXiv preprint arXiv:2407.18060*, 2023.

[62] M. Zhou, J. Scott, and B. Chaudhary. Enhancing Diagnostic Accuracy in Prostate Cancer with Integrated Radiomic Features: a Deep Learning Approach. *Journal of Medical Imaging and Radiation Oncology*, 66(3):342–356, 2022.

[63] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[64] Alfréd Rényi. On Measures of Entropy and Information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561, 1961.

[65] Xuelei Zhou and et al. Radiomics in Prostate Cancer: a Systematic Review and Meta-analysis. *Scientific Reports*, 10(1):970, 2020.

[66] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[67] Chris Ding and Hanchuan Peng. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational Biology*, 3(02):185–205, 2005.

[68] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature Selection Based on Mutual Information: Criteria of Max-dependency , Max-relevance , and Min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[69] M Ahmed and et al. Enhancing Cross-vendor Reproducibility of AI Models for Prostate Cancer Detection: a Radiomics Approach. *Journal of Medical Imaging*, 10(4):044501, 2023.

[70] Tiffany K Bell, Kate J Godfrey, Ashley L Ware, Keith Owen Yeates, and Ashley D Harris. Harmonization of Multi-site Mrs Data with Combat. *NeuroImage*, 257:119330, 2022.

[71] Fahimeh Mirakhori and Sarfaraz K Niazi. Harnessing the AI / ML in Drug and Biological Products Discovery and Development: the Regulatory Perspective. *Pharmaceuticals*, 18(1):47, 2025.

Original Publications

Chaudhary, Jatin Kumar and Liu, Jiaqing and Skön, Jukka-Pekka and Chen, Yen Wie and Kanth, Rajeev Kumar and Heikkonen, Jukka
Optimization of Silicon Tandem Solar Cells Using Artificial Neural Networks

Research and Development in Intelligent Systems XXXVI, pp. 392–403, 2019.

I

Optimization of Silicon Tandem Solar Cells Using Artificial Neural Networks

Jatin Kumar Chaudhary[1(✉)], Jiaqing Liu[2], Jukka-Pekka Skön[3], Yen Wie Chen[2], Rajeev Kumar Kanth[3], and Jukka Heikkonen[1]

[1] University of Turku, Turku, Finland
jatinkchaudhary@gmail.com
[2] Ritsumeikan University, Kyoto, Japan
[3] Savonia University of Applied Sciences, Kuopio, Finland

Abstract. The demand for photovoltaic cells has been increasing exponentially in the past few years because of its potential for generating clean electricity. Yet, due to low efficiency, this technology has not demonstrated complete reliability and poses tremendous amount of constraints even after the possibility of substantial power outputs. The concept of multi-junction solar cell has provided partial solution to this problem. Since the multi-junction solar cell was developed, its optimization has posed a great challenge for the entire community. The present study has been conducted on Si tandem cell, which is a two-junction three-layered solar cell. Silicon (Si) tandem cell was one of the initial developments in the domain of multi-junction solar cells and is most commercially fabricated photovoltaic cell. In this paper, the optimization challenge of multi-junction solar cells has been attempted with the use of Artificial Neural Network (ANN) technique. Artificial Neural Network was trained by using Bayesian Regularization algorithm, and used. Input parameters were taken as spectral power density, temperature and thickness of the layers of cells. Voltage of the cell was kept as a biasing input, and the output parameter was taken to be current density. I-V characteristics were plotted which was further used to calculate the open-circuit voltage (Voc), Fill Factor of the cell (FF), short circuit current density (Jsc) and Maximum Power Point (MPP). The output generated by the trained model of ANN has been compared with the values generated by more than a million iteration of the solar cell model. The implementation of this algorithm on any model of the multi-junction solar cell can lead to the development of highly efficient solar cells. Thus, with due consideration of physical constraints of the environment where it is to be installed; maximum amount efficiency can be achieved.

Keywords: Artificial Neural Network · Multijunction photovoltaic cell · Optimization

1 Introduction

The development of the photovoltaic cell is one of the most significant steps that mankind has taken towards research and practice on clean energy. The invention is a distinctive step towards the reduction in carbon emissions being rendered by

M. Bramer and M. Petridis (Eds.): SGAI-AI 2019, LNAI 11927, pp. 392–403, 2019.
https://doi.org/10.1007/978-3-030-34885-4_30

continually pronounced increments in carbon-based fossil fuel consumption by the post-modern economies. Apart from its potential to render zero carbon emissions, solar cell technology has also gained significant attention because of the economic benefits it posits [1]. At present, research has culminated in successful development and fabrication of the third-generation photovoltaic cells which are multi-junction thin-film solar cells. Yet, it is argued that the potential of this invention remains largely untapped. Also, motivation for growth in the invention, as well as further development of this field remains equivalent to what it was decades ago [2].

The efficiency of a solar cell, i.e., its energy conversion ratio, because of weakly optimized and modestly efficient cells has emerged as a significant challenge during the development of this technology. Furthermore, the economic output and adoption of the multi-junction solar cell remains significantly less than its theoretically calculated potential [3]. While it can be said that the efficiency of solar cells has been increased through the implementation of modern fabrication techniques. There is a need to examine methods for further optimization of parameters to improve the outcomes and hence, bring out the best possible efficiency ratio in a particular solar cell. One such method with high potential to address this problem is the integration of Artificial Neural Networks (ANN) in the field of solar cells.

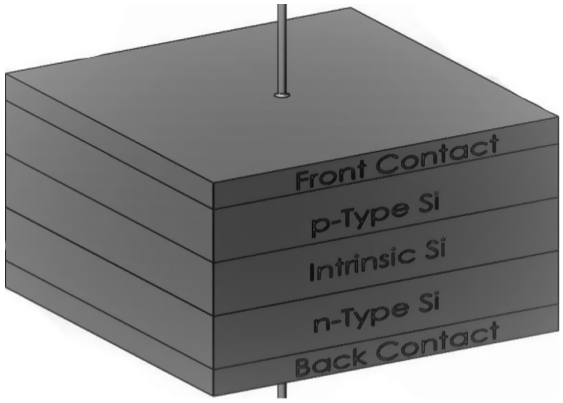


Fig. 1. Si tandem cells used for the development of datasets which has been used to train the ANN.

ANN has been successfully applied in prior studies of different parameters of solar cells, especially for tracking maximum power point prediction - an extreme measure of efficiency [4]. State-of-art research in the field is currently focused on training multiple neural networks by different techniques to predict the most suited neural network model for a given solar cell [4]. The research community is rigorously working towards the development of an ANN which can optimize a multi-junction photovoltaic technology.

This study focuses on developing an ANN model by training data taken by multiple iterations of a silicon tandem cell, which, as per our knowledge, is one of the most commercially fabricated multi-junction solar cells [5]. After training the ANN model by different training algorithms, Bayesian Regularization algorithm was found to give the best result with the least error, and the resultant output has been presented in this paper. The values extracted from the neural network model have been used to report the parameters of optimized model and further, to get the IV curves. There has been a thorough comparison between the optimized values attained by iterations on SCAPS (Solar Cell Capacitance Simulator) and the optimized values achieved from the ANN. Figure 1 shows the structure of the Si tandem cells on which the study has been done. This study can be taken further for utilization in various geographical regions, post - consideration of spectral power density and temperature of the location. Hence, a cell may be modelled according to different instalment sites to get the most efficient and optimized output for a particular instalment site.

2 Previous Works

A considerable amount of research has been conducted to optimize several performance parameters of photovoltaic systems in different contexts with the efficient use of an ANN. For instance, maximum power point tracking algorithms based on ANN may force photovoltaic modules to operate at their maximum power points for all environmental conditions [11]. Further, an ANN-based algorithm may correctly track the maximum power point even under abrupt changes in solar irradiance and improve the dynamic performance across the DC capacitor in the power converter that serves as interphase to connect photovoltaic power plants into the AC grid [7]. A summary of critical studies related to the use of ANN in optimizing the performance of photovoltaic systems in chronological order is presented in Table 1. However, the existing literature suggests that almost all of the experiments have been conducted with single-junction solar cells. Therefore, there is a research gap on optimizing performances of multi-junction solar cells, and the present study aims to address the same.

3 Dataset Description

SCAPS is a one-dimensional solar cell simulation program developed at the Department of Electronics and Information Systems (ELIS), University of Gent, Belgium [5]. The photovoltaic cell in Fig. 1 was iterated by simulations for various spectral densities, temperature and thickness of the three layers and hence, the I-V characteristics along with J_{SC}, V_{OC}, and FF of the cell were noted. In total, 5143 iterations were run and 61,716 values were calculated that were used to train the neural network model for generating the dataset for the Neural Network model hence developed. Further, this dataset was split into Training (70%), Testing (10%) and Validation (20%) so as to get maximum output from the modelled ANN. The dataset was generated by varying all the input parameters and calculating the current density values with voltage as a biasing unit.

Table 1. Summary of important studies.

Author(s) (year)	Study objective	Study findings
Kalogirou [9]	Use of artificial intelligence methods like neural networks and genetic algorithms in optimizing a solar-energy system	A system is modelled using a TRNSYS computer program where weather conditions are embedded to the input data used to train the model. Such methods pioneered the optimization of complicated solar-energy systems
Karatepe, Boztepe and Colak [10]	An application of artificial neural networks to photovoltaic module modelling	The dependence on environmental factors of the circuit parameters involves a set of nonlinear relationships that are difficult to express by analytical equations. However, a neural network may overcome the difficulty
Bae, Jeon, Kim, Kim, Kim, Han and May [6]	Techniques for optimizing processes in cascaded solar cell fabrication following neural networks and genetic programming modelling	The five variables, namely, texturing time, amount of nitrogen, DI water, diffusion time, and temperature are key to the recipe for solar cell fabrication. Repeated applications of particle swarm optimization yielded process conditions with smaller variations, and, greater consistency in recipe generation
Rai, Kaushika, Singh and Agarwal [12]	An artificial neural network based maximum power point tracking controller to predict maximum power voltage and maximum power current under varied atmospheric and load conditions	A model for the energy generation by a photovoltaic array has been developed to capture the effect of solar irradiance, atmospheric temperature, wind speed and variability of the load in the circuit. Its maximum power point tracking performance excels over the conventional PID controller and avoids the tuning of controller parameters
Subiyanto, Mohamed and Hannan [13]	A method for maximum power point tracking of a photovoltaic module by using the Hopfield neural network optimized fuzzy logic controller	Simulation and experimental results show that the method proposed in the study is more robust and accurate compared to the conventional methods. Further, this method successfully tracks the global maximum power point of a photovoltaic energy harvesting system

(*continued*)

Table 1. (*continued*)

Author(s) (year)	Study objective	Study findings
Chen, Gooi and Wang [8]	Use of fuzzy and neural networks to forecast solar radiation accurately at different weather conditions	The mean absolute percentage error produced by this technique is much smaller compared to that of the other methods when used in grid-connected photovoltaic systems
Yuan, Xiang and He [14]	A mutative-scale parallel chaos optimization algorithm using crossover and merging operation to optimize photovoltaic system performance	This technique outperforms other meta-heuristic algorithms commonly deployed for extracting different parameters of solar cell models, such as double diode, single diode, and photovoltaic module, among others

4 The Architecture of an Artificial Neural Network

MATLAB® (2019a) was used to perform ANN modelling. The Neural Network Toolbox of MATLAB has been used for the fitting of the curve and creation of a successful model of the network [15].

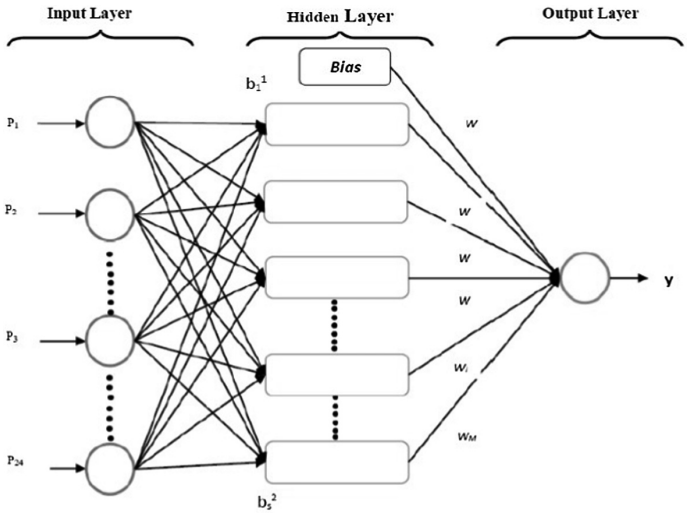


Fig. 2. Basic architecture of an Artificial Neural Network.

Multilayer Perceptron (MLP) Feed Forward Fully Connected Neural Network has been modelled for optimization of parameters. In MLP networks, there is an input

layer, an output layer, and hidden layers. The hidden layers are associated with the weights which are calculated during the training of the model. Figure 2, shows the basic architecture of an ANN. Multilayer Perceptron network was trained with five inputs parameters (Spectral Power Density, Temperature, the thickness of three layers of the cell), one biasing parameter (Voltage), one hidden layer with 40 neurons and one output parameter (Current Density). In this study, series and shunt resistances have been taken as zero, i.e. ideal condition of a solar cell has been considered.

The Tansig function (Hyperbolic tangent sigmoid transfer function) has been used as the activation function for the neural network. This hyperbolic tangent transfer function is related to a bipolar sigmoid which has an output in ranging from {−1 to +1}. Tansig function is given by the following equation:

$$\mathrm{n} = \frac{2}{1 + e^{-2n}} \tag{1}$$

The Purelin function has been used in the network as the transfer function of the output layer, which as a linear transfer function denoted by:

$$\Psi(\mathrm{n}) = \mathrm{n} \tag{2}$$

ANN is trained to minimize error between the target to be achieved and the input parameters. Further, to get the desired ANN i.e. with least error and best fir parameter, the error is minimized by adjusting the weights and biases which has been computed in the previous cycle of training. By dynamic updating of weights and biases, a reliable network i.e. least error is obtained. The ANN with minimum error is hence used to predict data for inputs. Objective function of the training algorithm of Neural Network is given by:

$$\mathrm{E} = \frac{1}{\mathrm{n}} \sum_{i=0}^{n} |y_i - x_i|^2 \tag{3}$$

where, E = Error of the Neural Network,

$$x = \text{Input data} = [x_1, x_2, x_3, \ldots\ldots, x_n]$$

$$y = \text{Target datasets} = [y_1, y_2, y_3, \ldots\ldots, y_n]$$

It can be easily noticed that this is the equation for the Mean Squared Error and hence, the minimization of this equation is the objective as that implies minimization of the error incorporated during the training process. The extensive use of regularization techniques is to ensure the freedom from the overfitting of data.

In particular, the Bayesian Regularization algorithm enhances the performance of ANN and makes it more reliable by minimizing the error. This is achieved by an algorithm which penalizes the sum of the squared errors and revises reassignment the weights dynamically according to the magnitude of the error that had occurred. This modified objective function of the basic ANN is considered to be a Bayesian Function

and the parameters i.e. the biases and the weights are taken to be a random variable. The updated objective function is:

$$F = \beta E_D + \alpha E_w \tag{4}$$

where, E_w = Sum of squares of the Network weight
E_D = Sum of squared errors
α, β = Objective function parameter or weight coefficient.

The error incorporated in the network depends on the variation of the weight coefficients which gets dynamically updated in an ANN. If $\alpha << \beta$, then the training algorithm will derive smaller error, whereas when $\alpha >> \beta$, training will tend to reduce weight size, at an expense of network error but would produce a smoother network response [16, 17].

The weight of the Bayesian Regularization algorithm is updated by the back-propagation technique which uses the following equation:

$$w_x = w_x - \alpha\left[\frac{\partial E}{\partial \omega}\right] \tag{5}$$

Where, w_x = Weight assigned to the parameter
α = Learning rate
$\left[\frac{\partial E}{\partial \omega}\right]$ = Derivative of error w.r.t. weight

The weights after the training of the ANN contain meaningful information, i.e. they convey the relationship between the input and the target, whereas before training the weights are just random values without any implacable meaning.

5 Results and Discussions

The network was modelled and trained as per previously discussed process. Multilayer Perceptron (MLP) Feed Forward Fully Connected Neural Network with one hidden layer consisting of 40 neurons was trained with Bayesian Regularization Algorithm on MATLAB (2019a) software on a system with memory of 16 GB. After training, the network outcomes were reported as per the information presented in Table 2.

Table 2. The Neural network training parameters and results.

Training Samples	3600
Validation Samples	1028
Testing Samples	514
Mean Square Error for Training	0.00313071
Mean Square Error for Test Samples	0.00377475
Regression Values for Training Samples	0.9998
Regression values for Test Samples	0.99978
Time Taken	30 s

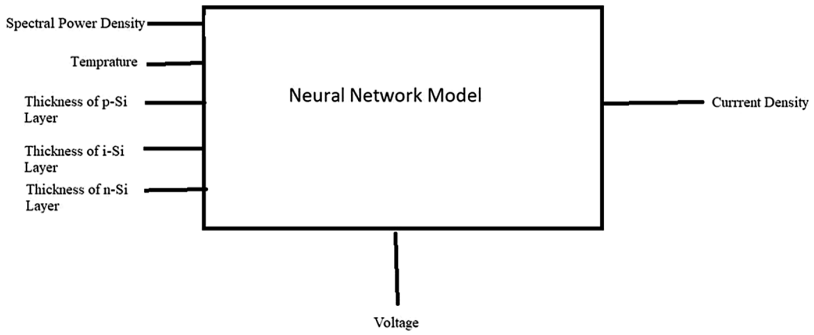


Fig. 3. Architecture of the modeled neural network with five input, one biasing and one output parameter.

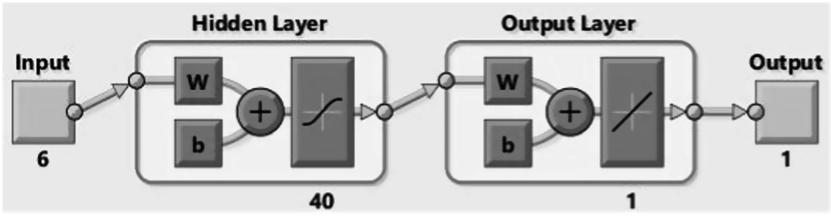


Fig. 4. The neural networks with 40 hidden neurons and an output layer with the Tansig and the Purelin function, respectively.

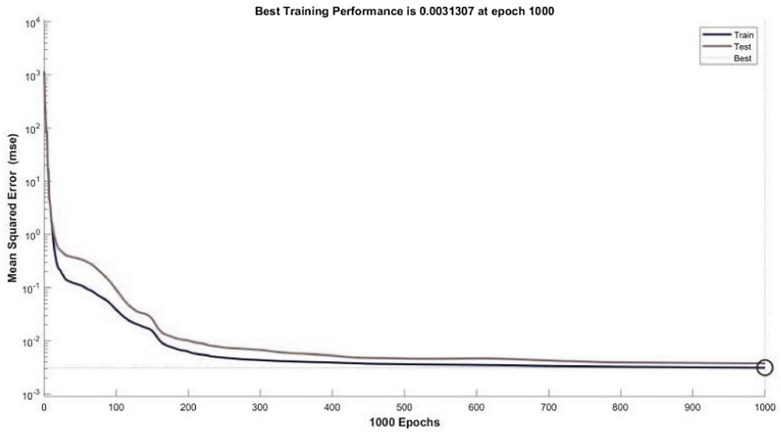


Fig. 5. Performance graph for neural network trained.

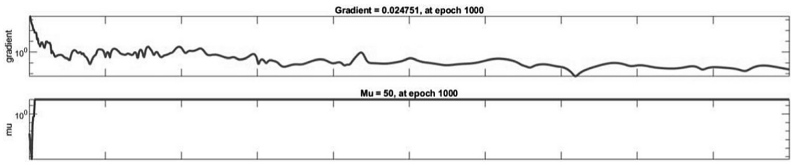


Fig. 6. Training state of neural network model.

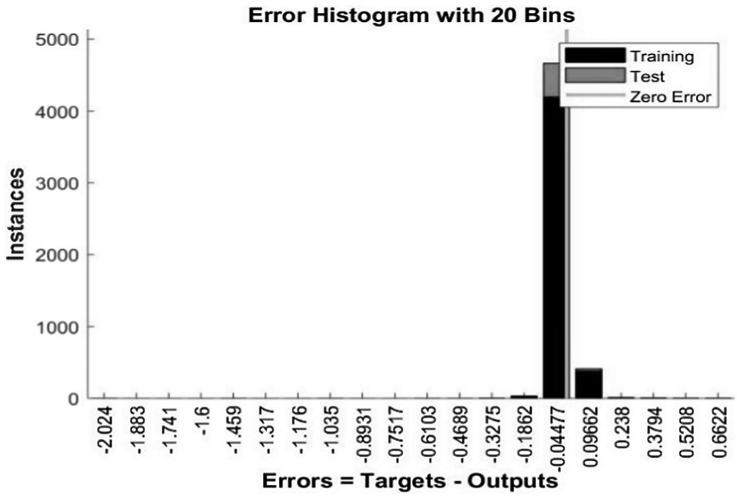


Fig. 7. Error histogram of the trained model.

Figures 3, 4, 5 and 6 show the modeled neural network architecture, structure, performance and training state. Figure 7 represents the regression curve of the model where it can be seen that a linear line passes through almost 90% of the data. Hence, the regression model has a value close to 1 and implies to be efficient enough to give a reliable data This also indicates that the relationship between the input parameters and the target values are relatable.

In order to get the optimized model through simulations, more than one million iterations were performed in SCAPS and the points for optimization obtained by simulation were noted. After the training of the ANN, points of optimization produced by the ANN was noted from the fit curve obtained by 'Plotfit' function. Table 3, shows the comparison of the optimized points obtained from the ANN model and computational experimentation performed on SCAPS.

It can be noted that the points obtained by both the experiments has very less quantitative difference. Analyzing both the methods of obtaining optimization points, we come to an inference that the ANN takes only 30 s to be trained to produce a result whereas SCAPS takes a relatively longer time and more computational expense.

Table 3. Points of optimization obtained from the trained model.

	SCAPS	ANN
Spectral Power Density (W m-2)	767.17	861.6
Temperature (K)	354.37	349.3
Thickness of p-layer (μm)	9.9678	9.45
Thickness of i-layer (μm)	1.897	1.61
Thickness of n-layer (μm)	6.4557	8.5

Using both the output points, a simulation of the Si tandem solar cell was performed on SCAPS and Fig. 8, shows the IV characteristic of the cell. It can be observed that the curves are optimized and which is also proven by various parameters on which efficiency is directly dependent. Table 4, shows the data obtained from the IV characteristics of the optimized cell from both the techniques.

Table 4. The efficiency measures of the photovoltaic cell whose parameters were calculated by SCAPS and ANN.

	Optimization points obtained from SCAPS	Optimization points obtained by ANN
Open Circuit Voltage (Voc) (V)	0.7804	0.794719
Short Circuit Current Density (mA cm-2)	11.318835	11.37040531
Fill Factor (%)	57.00	58.8717
Maximum voltage point (V_{MPP})	0.543262	0.567910

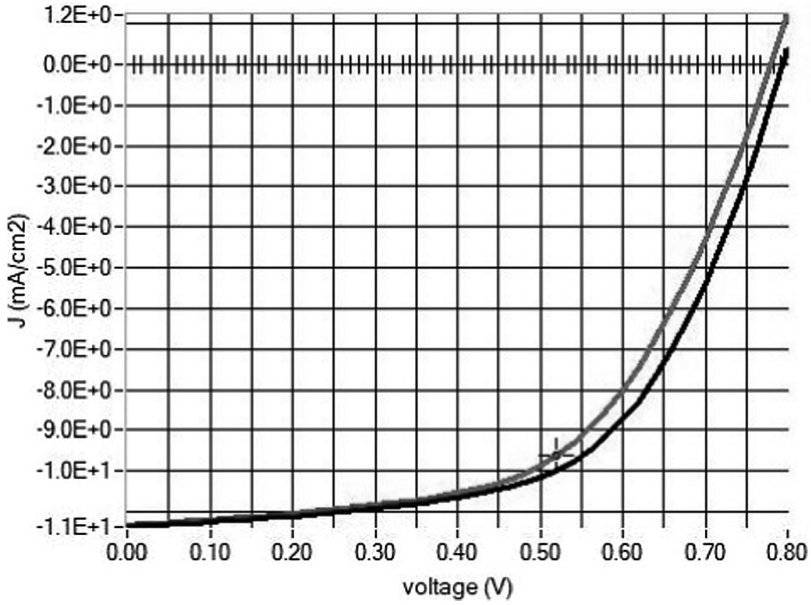


Fig. 8. IV characteristic of the optimized solar cell as from the output of SCAPS and neural network model. (Left (Grey) = output from SCAPS and Right (Black) = output from ANN).

6 Conclusion

This study was based on the significant need for the optimization of a multi-junction solar cell. In this research, modeling was done on the data obtained for a Si tandem cell using Artificial Neural Network. A network with 40 hidden neurons was developed on MATLAB which was trained using Bayesian Regularization algorithm through 61,716 values obtained by iterations done on the cell. This trained network was used to predict values of the input parameters to generate the most optimum model of the solar cell. This model was compared with optimization values obtained by SCAPS. These values were tested on the Si tandem cells and the obtained I-V characteristic proved that the algorithm worked well. It also implicates the efficient capability of predicting highly optimized multi-junction solar cells.

References

1. Roessner, J.D.: Government-industry relationships in technology commercialization: the case of photovoltaics. Sol. Cells **5**(2), 101–134 (1982)
2. Third Generation Photovoltaics (2012)
3. Gul, M., Kotak, Y., Muneer, T.: Review on recent trend of solar photovoltaic technology. Energy Explor. Exploit. **34**(4), 485–526 (2016)

4. Ramaprabha, R., Gothandaraman, V., Kanimozhi, K., Divya, R., Mathur, B.L.: Maximum power point tracking using GA-optimized artificial neural network for solar PV system. In: 2011 1st International Conference on Electrical Energy Systems, pp. 264–268. IEEE, January 2011
5. Burgelman, M., Nollet, P., Degrave, S.: Modelling polycrystalline semiconductor solar cells. Thin Solid Films **361**, 527–532 (2000)
6. Bae, H., et al.: Optimization of silicon solar cell fabrication based on neural network and genetic programming modeling. Soft. Comput. **14**(2), 161–169 (2010)
7. Carrasco, M., Mancilla-David, F., Fulginei, F.R., Laudani, A., Salvini, A.: A neural networks-based maximum power point tracker with improved dynamics for variable dc-link grid-connected photovoltaic power plants. Int. J. Appl. Electro Magn. Mech. **43**(1–2), 127–135 (2013)
8. Chen, S.X., Gooi, H.B., Wang, M.Q.: Solar radiation forecast based on fuzzy logic and neural networks. Renew. Energy **60**, 195–201 (2013)
9. Kalogirou, S.A.: Optimization of solar systems using artificial neural-networks and genetic algorithms. Appl. Energy **77**(4), 383–405 (2004)
10. Karatepe, E., Boztepe, M., Colak, M.: Neural network based solar cell model. Energy Convers. Manag. **47**(9–10), 1159–1178 (2006)
11. Kulaksiz, A.A., Akkaya, R.: Training data optimization for ANNs using genetic algorithms to enhance MPPT efficiency of a stand-alone PV system. Turk. J. Electr. Eng. Comput. Sci. **20**(2), 241–254 (2012)
12. Rai, A.K., Kaushika, N.D., Singh, B., Agarwal, N.: Simulation model of ANN based maximum power point tracking controller for solar PV system. Sol. Energy Mater. Sol. Cells **95**(2), 773–778 (2011)
13. Subiyanto, S., Mohamed, A., Hannan, M.A.: Intelligent maximum power point tracking for PV system using hopfield neural network optimized fuzzy logic controller. Energy Build. **51**, 29–38 (2012)
14. Yuan, X., Xiang, Y., He, Y.: Parameter extraction of solar cell models using mutative-scale parallel chaos optimization algorithm. Sol. Energy **108**, 238–251 (2014)
15. MATLAB and Neural Network Toolbox Release: The Math-Works, Inc., Natick, Massachusetts, United States (2018b)
16. San, O., Maulik, R.: Neural network closures for nonlinear model order reduction. Adv. Comput. Math. (2018). https://doi.org/10.1007/s10444-018-9590-z
17. Dan Foresee, F., Hagan, M.T.: Gauss-Newton approximation to Bayesian learning. In: Proceedings of International Conference on Neural Networks (ICNN 1997) (n.d.). https://doi.org/10.1109/icnn.1997.614194

Chaudhary, Jatin and Bhattacharya, Swastik and Heikkonen, Jukka and Kanth, Rajeev
Prediction of Electron Band Gap of A2XY6 Perovskite Compounds using Machine Learning

IEEE 49th Photovoltaics Specialists Conference (PVSC),1173–1176, 2022.

II

Prediction of Electron Band Gap of A_2XY_6 Perovskite Compounds using Machine Learning

1st Jatin Chaudhary
Department of Computing
University of Turku
Turku, Finland
jatin.chaudhary@utu.fi

1st Swastik Bhattacharya
Department of Electrical, Computer and Energy Engineering
University of Colorado
Boulder, USA
Swastik.Bhattacharya@colorado.edu

2nd Jukka Heikkonen
Department of Computing
University of Turku
Turku, Finland

3rd Rajeev Kanth
School of Information Technology
Savonia University of Applied Sciences
Kuopio, Finland

2022 IEEE 49th Photovoltaics Specialists Conference (PVSC) | 978-1-7281-6117-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/PVSC48317.2022.9938521

***Abstract*—Increasing population and industrialization haveled to an uptick in energy requirements. Many traditional energy sources are not anymore attractive due to climate change, instead, the interest has turned to power generation from renewable sources, such as wind energy, hydro-power, and solar energy. The wide availability of sunlight and simplicity in converting sunlight to electricity has led to the search for synthesized semiconductors that give high efficiency in this conversion. A family of such semiconductors attains the perovskite structure, the most established being Methyl Ammonium Lead Iodide. The shortcomings of this compound include lead poisoning, motivating the search for perovskite structures that have low electron band-gap and are stable. A family of such perovskite structures is compounds that attain an A_2XY_6 type structure. This paper demonstrates some methods that can be used to calculate the electron band-gap of such compounds. The metrics found from Support Vector Machine Regression and Random Forest Regression are compared and analyzed to propose a scalable model for predicting electron band-gap.**

***Index Terms*—A_2XY_6 Perovskite Compounds, Band Gap Calculation, Support Vector Machine, Random Forest**

I. Introduction

The shift of energy dependencies from traditional origins to renewables has been prominently taking place in the past decade. Among all, solar Energy has been one of the most promising source of renewable energy [1]. In 2020-2021, solar energy has been second most eminent renewable source [2]. Researches to improve the efficiency of photovoltaic(PV) cells has been an imperative task looking over the potential PVs possess [3]. Over the years, the perovskite technology has accelerated the optimization of PV cell due to its stable nature and high potential of increased power conversion efficiency [4]. Perovskite Cells have definitive crystal structure as ABX_3 (where, X= oxygen, hallide). A_2XY_6 (A= K, Cs, Rb, Tl; X= tetravalent cation, Y= F, Cl, Br, I) Perovskite compunds have been identified to have incredible semiconductor properties along with exclusive optical properties making it suitable for opto-electronic devices like PV cells [5].

Designing of highly efficient PV cells have been a notable state-of-the-art research in the field and the inter-linkage to the perovskite technology with PV have been very assuring [6] [7]. The increasing investigation of perovskite compounds are bringing an evolution to the PV industry [8]. Band gap energies are a crucial consideration when the PV cells are designed as a wide band gap energies may make the compound unsuitable for PV applications [9]. Calculation of the bandgaps for different compounds is an cumbersome process as it requires optical diffuse reflectance measurements followed by computations using Kubelka–Munk equation [10]. Studies have shown that direct band gaps can vary with the experimental structure as band gap is susceptible to functional group employed for structure optimization [11].

Keeping Volonakis et al., 2017's study in consideration, the authors of this paper decided to study the dependence of bandgap upon the experimental structure of the compound and further, a method to predict the bandgap based on the experimental structure. This has been done so as to predict the bandgap of newly simulated or developed A_2XY_6 Perovskite Compounds. In this paper, we have presented an efficient method for prediction of band gap of A_2XY_6 Pervoksite Compounds using Support Vector Machine (with Radial basis Kernel and Linear Kernel) and Random Forest [12] [13]. This study has been performed in order to present a solution to the band gap energy for prediction of upcoming compounds which are in the primitive stage of designing or manufacturing. Physical parameters including ionic radii, electronegativities, Miller Indices, formation energy, lattice index and lattice constants are used for the training of the model. The data used for the training of the model is taken from the literature available upon the experimentation.

The paper consists of methodology in section II, and results of the model trained and it's discussion in section III, and conclusion and future scope of the work in section IV.

II. Methodology

A. Datasets

Data used for the training of models consists of experimental output of eighty nine A_2XY_6 perovskite compounds with lattice constant of 8.109 to 11.790 Å [14]. Seventy one compounds as data points are taken from the [14], whereas the rest are the repeated compounds which exist in more than one structure, hence having different Lattice Constants and Miller Indices. The value of lattice constants has been taken in normal room temperature and pressure conditions. Compounds which qualified the condition A_2XY_6 perovskite compounds with lattice constant of 8.109 Å to 11.790 Å but had unstable structural values have not been considered as a part of the dataset.

B. Models and model performance evaluation

For the purpose of this study, due to a simpler complexity of the dataset, Random Forest (RF) Regression and Support Vector Machines (SVMs) with linear and radial-basis function kernels. The models were evaluated using Leave-One Out (LOO) cross-validation approach, having divided the data into 5 batches and cross-validating the results after leaving one of the batches.

A Random Forest (RF), as defined by Breiman et al. [15] is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$ where the Θ_k are independent identically distributed (i.i.d.) random vectors and each tree casts a unit vote for the most popular class at input **x**. RF for regression problems are created by growing trees depending on a random vector, such that the tree estimator h(**x**,Θ) takes on numerical values, and not class labels [15]. The output values are numerical and it is assumed that the training set is independently drawn from the distribution of the random vector Y, **X**. The mean-squared generalization error for any numerical predictor $h(\mathbf{x})$ is:

$$E_{\mathbf{X},Y}(Y - h(\mathbf{x}))^2 \tag{1}$$

SVMs are used in both classification and regression problems. The goal of SVMs in a regression problem is to find a function $f(x)$ that can provide predictions under a margin of error ε. The function $f(x)$ can be expressed as:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b \tag{2}$$

where $\Phi(\mathbf{x})$ is the mapping result into the input space, $\mathbf{w}$ is the weight matrix, and b is the bias vector. The weight and bias are trained by minimizing the risk function:

$$R = min \frac{1}{2}||\mathbf{w}||_2^2 + C\frac{1}{l}L_\varepsilon(y, f(\mathbf{x})) \tag{3}$$

Regression problems that are of non-linear nature can be solved by using non-linear functions. For any input vectors $\mathbf{x}$ and $\mathbf{z}$, a kernel function must satisfy the condition:

$$k(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^T \Phi(\mathbf{z}) \tag{4}$$

In this study, the regression problem has been attempted by using linear and radial-basis function kernel. A linear kernel is expressed as:

$$k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z} \tag{5}$$

A radial-basis function kernel is defined as:

$$k(\mathbf{x}, \mathbf{z}) = exp(\frac{1}{2\sigma^2}||\mathbf{z} - \mathbf{x}||_2^2) \tag{6}$$

For the purpose of this study, the number of tree estimators in the RF regressor is assigned as 100. For the SVRs, for both kernels, the value of C in equation (3) for risk minimization is set to 100, and the error margin ε is set to 0.1. Rest of the hyper-parameters are assigned the default values as per the scikit-learn package of Python [16]. The model was trained using the experimental data discussed above.

TABLE I
Result Matrix of the Models Trained.

	RMSE	***Relative RMSE***
SVM (Linear Kernel)	2.20 eV	0.31
SVM (RBF Kernel)	2.30 eV	0.32
Random Forest	2.33 eV	0.32

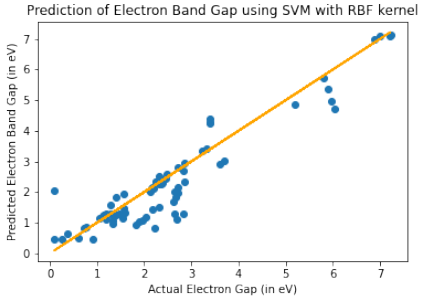


Fig. 1. Predicted vs Actual Bandgap graph of Radial Basis Function SVM.

III. Results and Discussion

The three models trained presented promising results. Table 1, shows the result matrix consisting of Root Mean Square Error(RMSE), and Relative RMSE for all the three models. Figure 1, depicts the predicted vs the actual band gap graph of the RBF SVM model trained. Similarly, figure 2 and 3 also depicts the predicted vs the actual band gap graph of the linear kernel SVM and random forest algorithm respectively. These graphs are an indication to the accuracy of the model. Furthermore, the error curves for the cross-validation of the SVM models were analyzed. These were obtained by performing a 5-fold Leave-One-Out cross validation over the models. For both the kernels in case of SVM, as shown in figures 4 and 5, the MSE converged to 1.25eV in case of linear

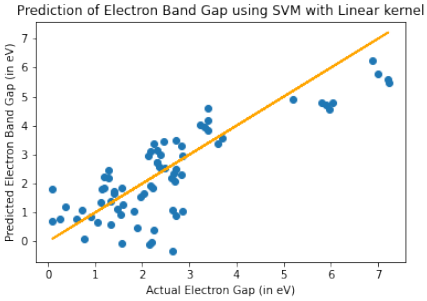


Fig. 2. Predicted vs Actual Bandgap graph of Linear SVM.

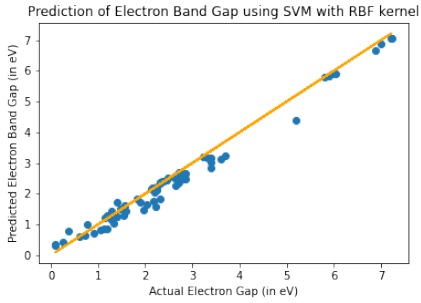


Fig. 3. Predicted vs Actual Bandgap graph of Random Forest.

kernel SVM, and 1.5eV in case of the RBF kernel SVM. The convergence of all the models affirm the optimization of the model. Figures 6, 7 and 8 depict scatter plots between the ground-truth and cross-validation predictions. It is observed that there is a better fitting on testing ground-truth in case of RF regression upon cross-validation as compared to both the SVM models. However, as stated in table 1, it is also seen that the overall accuracy in terms of RRMSE is better in case of both the SVM models.

The calculation of the bandgap of A_2XY_6 perovskite compounds requires the computation over first principles and experimental results [11]. This study has been focused upon developing a model which can predict the band gap value using the physical properties (input features) of stable double perovskite material (A_2XY_6). Authors' focus has been towards developing a model which can be used to predict the bandgap of the newly simulated A_2XY_6 Perovskite Compounds. This method of bandgap prediction has been developed to bypass the computationally expensive calculations of bandgap, for further estimation of the compound's practical viability. Our focus has been towards building a model which can predict bandgap values which can be made into a scalable solution to avoid the current computationally intensive models meant for this problem.

IV. Conclusion and Future Scopes

Our paper presents novel output in terms of prediction of band gap in a computationally economically way towards bandgap prediction of A_2XY_6 Perovskite Compounds. Among the models we have trained, the random forest model has given promising results towards the predictor function. The support vector machine with radial basis function and linear function has been trained, and Support Vector Machine with Radial Basis Function has given better results considering the potential intervention of noise signals while training. This paper can be extended with the addition of latest compounds to the dataset and optimizing the existing model with the feedback of more validation methods. This study has presented

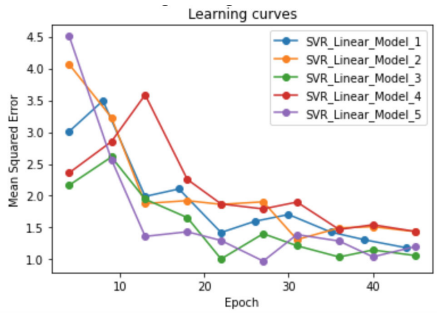


Fig. 4. Learning Curve for SVM Linear Kernel. Mean Square Error was calculated to be 1.25eV.

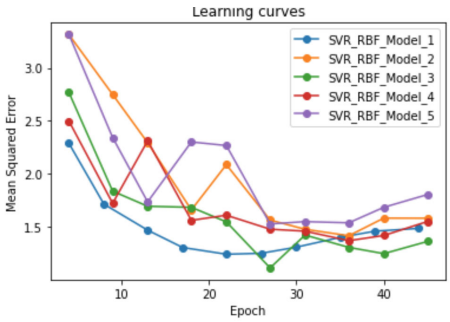


Fig. 5. Learning Curve for SVM RBF Kernel. Mean Square Error was calculated to be 1.5eV.

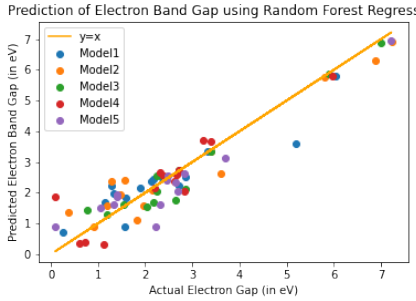


Fig. 6. Cross Validation graph of Random Forest.

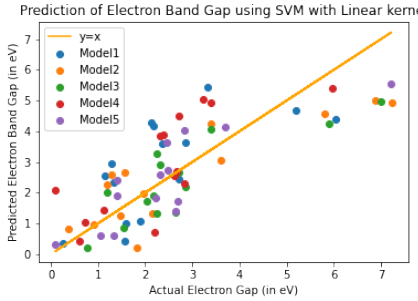


Fig. 7. Cross Validation graph of Support Vector Machine with Linear Kernel.

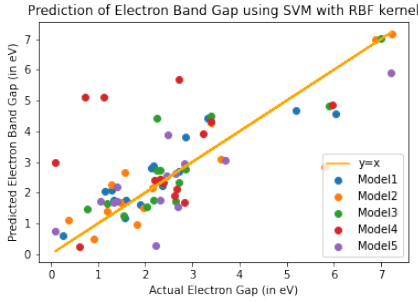


Fig. 8. Cross Validation graph of Support Vector Machine with RBF Kernel.

a preliminary study in the field of bandgap prediction of A_2XY_6 Perovskite Compounds and hence, possesses huge opportunities towards extending this study. Such efforts can also be applied towards predicting bandgap for perovskite compounds that have a different stoichiometry and elements present in their structure which may not necessarily form a perovskite of A_2XY_6 nature.

References

[1] N. Kannan and D. Vakeesan, "Solar energy for future world:-a review," *Renewable and Sustainable Energy Reviews*, vol. 62, pp. 1092–1105, 2016.

[2] Iea, "Renewable electricity generation increase by technology, 2019-2020 and 2020-2021 –nbsp;charts – data amp; statistics." [Online]. Available: https://www.iea.org/data-and-statistics/charts/renewable-electricity-generation-increase-by-technology-2019-2020-and-2020-2021

[3] J. K. Chaudhary, R. Kanth, J.-P. Skön, and J. Heikkonen, "Analysis and enhancement of quantum efficiency for multi-junction solar cell," in *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*, 2019, pp. 0210–0214.

[4] N.-G. Park, "Perovskite solar cells: an emerging photovoltaic technology," *Materials Today*, vol. 18, no. 2, pp. 65–72, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369702114002570

[5] S. Chadli, A. Bekhti Siad, M. Baira, M. Siad, A. Allouche, and A. Reguig, "Physical properties of double perovskites rb2xcl6 (x= sn, te, zr): Competitive candidates for renewable energy devices," *Solid State Communications*, vol. 342, p. 114633, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038109821004105

[6] J. Chaudhary, R. Kanth, and J. Heikkonen, "Performance analysis of back surface field (bsf) effects in multijunction photovoltaic cell," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, 2020, pp. 1207–1211.

[7] M. I. H. Ansari, A. Qurashi, and M. K. Nazeeruddin, "Frontiers, opportunities, and challenges in perovskite solar cells: A critical review," *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, vol. 35, pp. 1–24, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389556717301144

[8] A. Mutalikdesai and S. K. Ramasesha, "Emerging solar technologies: Perovskite solar cell," *Resonance*, vol. 22, no. 11, pp. 1061–1083, 2017.

[9] R. Nechache, C. Harnagea, S. Li, L. Cardenas, W. Huang, J. Chakrabartty, and F. Rosei, "Bandgap tuning of multiferroic oxide solar cells," *Nature Photonics*, vol. 9, no. 1, pp. 61–67, 2015.

[10] I. Chung, J.-H. Song, J. Im, J. Androulakis, C. D. Malliakas, H. Li, A. J. Freeman, J. T. Kenney, and M. G. Kanatzidis, "Cssni3: semiconductor or metal? high electrical conductivity and strong near-infrared photoluminescence from a single material. high hole mobility and phase-transitions," *Journal of the American Chemical Society*, vol. 134, no. 20, pp. 8579–8587, 2012.

[11] G. Volonakis, A. A. Haghighirad, R. L. Milot, W. H. Sio, M. R. Filip, B. Wenger, M. B. Johnston, L. M. Herz, H. J. Snaith, and F. Giustino, "Cs2inagcl6: a new lead-free halide double perovskite with direct band gap," *The journal of physical chemistry letters*, vol. 8, no. 4, pp. 772–778, 2017.

[12] G. L. Prajapati and A. Patle, "On performing classification using svm with radial basis and polynomial kernel functions," in *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, 2010, pp. 512–515.

[13] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[14] Y. Zhang and X. Xu, "Machine learning lattice constants from ionic radii and electronegativities for cubic perovskite a2xy 6 compounds," *Physics and Chemistry of Minerals*, vol. 47, no. 9, pp. 1–15, 2020.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

Chaudhary, Jatin and Jambor, Ivan and Boström, Peter and Saunavaara, Jani and Etala, Otto and Taimen, Pekka and Aronen, Hannu and Heikkonen, Jukka and Kanth, Rajeev and Merisaari, Harri

Foundational AI and Radiomics: Improving Reproducibility in Clinical Decision Support Systems for Prostate MRI

Submitted to Nature Cancer

Foundational AI and Radiomics: Improving Reproducibility in Clinical Decision Support Systems for Prostate MRI

Jatin Chaudhary[1,*], Ivan Jambor[2], Peter Boström[3], Jani Saunavaara[4], Otto Ettala[3], Pekka Taimen[3], Hannu Aronen[2], Jukka Heikkonen[1], Rajeev Kanth[5], and Harri Merisaari[2]

[1]Department of Computing, University of Turku, Turku, Finland
[2]Department of Diagnostic Radiology, University of Turku, Turku, Finland
[3]Department of Urology, University of Turku, Turku, Finland
[4]Department of Medical Physics, Turku University Hospital, Turku, Finland
[5]Savonia University of Applied Sciences, Kuopio, Finland
[*]Jatin Chaudhary (jkchau@utu.fi)

ABSTRACT

Clinical decision-making for prostate cancer (PCa) remains limited by expert interpretation and heterogeneous imaging protocols. We developed a foundational Artificial Intelligence model using radiomics-derived features from multicenter prostate MRI data to enhance PCa detection. This retrospective study analyzed 1,152 patients who underwent T2-weighted and diffusion-weighted imaging (DWI) on 1.5T and 3T MRI scanners, with lesions manually delineated and matched with histopathology. The model was trained to classify significant vs non-significant cancer. A Vision Transformer (ViT)-based model achieved an AUC of 0.86 on the training dataset. Fine-tuning on site specific homogeneous datasets of only 60 patients per site improved performance, reaching an AUC of above 0.90 in external validation. This fine-tuning mechanism was validated across multiple datasets, demonstrating reproducibility and clinical scalability. To enhance transparency, we employed Explainable AI (xAI) to elucidate key radiomics features influencing predictions. Our findings demonstrate that a large foundational model can be effectively fine-tuned with minimal data, ensuring robust deployment across diverse clinical sites.

Introduction

Prostate cancer (PCa) remains one of the most prevalent and lethal malignancies among men worldwide, representing a significant burden on public health systems across the globe. According to statistics for 2024, the mortality rates associated with prostate cancer continue to be alarming, with the United States alone witnessing over 35,000 deaths annually, while Europe predicts 72,059 deaths and Asia report similar concerning figures[1,2]. Globally, it is estimated that prostate cancer accounts for over 375,000 deaths each year, underscoring the urgent need for improved diagnostic and therapeutic strategies. The current diagnostic protocol for prostate cancer primarily relies on prostate-specific antigen (PSA) testing, digital rectal examinations (DRE), and, in cases of elevated risk, multiparametric magnetic resonance imaging (mpMRI)[3]. In scenarios of scanning the patient with mpMRI, Prostate Imaging Reporting and Data System (PIRADS) is instrumental in standardizing the interpretation of mpMRI findings, facilitating the classification of lesions based on their likelihood of being clinically significant cancer[4,5]. PIRADS scores range from 1 to 5, with higher scores indicating a greater probability of malignancy. Further diagnostic refinement is achieved through the assessment of Gleason Grade Groups (GGGs), which are based on histopathological evaluations of prostate biopsy specimens[6]. The GGG system classifies prostate cancers into five groups, with GGG 1 representing the least aggressive form of the disease and GGG 5 indicating the most aggressive and poorly differentiated tumors indicating to current or near future metastases[7]. The combination of PIRADS and GGGs provides a comprehensive framework for risk stratification, guiding clinical decision-making in the diagnosis and prognosis of prostate cancer. Despite these advances, the complexity of prostate cancer diagnosis necessitates the integration of state-of-the-art methodologies, such as image processing, artificial intelligence etc., to enhance the accuracy and prognostic capabilities of existing diagnostic paradigms.

Alongside the classical way of diagnosis by Radiologists, usage of artificial intelligence alongside the human expertise towards diagnosis has come a long way. The application of Artificial Intelligence (AI) in the field of prostate cancer diagnosis and prognosis has leveraged substantial interest, as demonstrated by a plethora of recent studies. Different techniques are being utilized for delination, segmentation and classification of tumour[8,9]. Sandeman et al. (2022) focused on the development

of a Convolutional Neural Network (CNN) to assist in cancer diagnosis, Gleason grading, and prognostic correlation in prostate cancer. Utilizing the Pannoramic 250 Flash III scanner (3DHistech, Budapest, Hungary), this study analyzed 4,221 biopsy slides and 750 patient samples, alongside prostatectomies from 126 patients. The AI models demonstrated high diagnostic performance, with cancer detection sensitivities and specificities reaching 98% and a Kappa coefficient of 0.96 when compared to the reference diagnosis. The study also highlighted a strong correlation between AI-based cancer detection and various adverse prognostic factors, such as extraprostatic extension and seminal vesicle involvement[10]. In a study by Leo (2021), a CNN-based approach was employed to correlate cribriform carcinoma presence with biochemical recurrence post-prostatectomy in 819 patients. The CNN, based on a U-Net architecture, utilized internal and external datasets to validate its prognostic capabilities, revealing a significant correlation between the calculated area of cribriform cancer and biochemical recurrence, as evidenced by a multivariate analysis, p=0.018[11]. Pantanowitz et al. (2020) sought to validate an AI-based algorithm for prostate cancer classification and the detection of clinically relevant features such as perineural invasion. The study included 910 cases with 4,677 slides scanned using the IntelliSite Scanner (Philips Digital Pathology Solutions) and the Aperio AT2 scanner (Leica). The AI model achieved an impressive ROC-AUC of 0.99 for cancer detection in the external validation set and 0.94 for distinguishing between low-grade and high-grade cancer, alongside a Kappa coefficient of 0.88 in comparison with light microscopy (LM)[12]. Nagpal et al. (2019) introduced a deep learning system aimed at Gleason grading on prostatectomy samples. The study involved 1,557 patients, with the slides being digitized using Aperio AT2 (Leica) and Hamamatsu scanners. The AI system, which was based on an InceptionV3 architecture, achieved a ROC-AUC of 0.95-0.96 and demonstrated a concordance index (c-index) of 0.69 with respect to biochemical recurrence and disease progression, aligning closely with the reference standard set by specialist pathologists[13]. Lucas (2019) proposed an automatic method for prostate cancer detection and Gleason grading on core needle biopsies (CNBs) using an Inception-v3 CNN. This study, which included 38 patients and 96 slides scanned with a Philips UltraFast scanner, reported an accuracy of 92% for cancer detection and 90% for differentiating between Gleason grades, with a Kappa coefficient of 0.7[14]. A study conducted by Masayuki Tsuneki (2022) focused on classifying transurethral resection of the prostate (TUR-P) slides into adenocarcinoma and benign lesions using deep learning models trained via transfer learning and weakly supervised learning. This analysis, which involved 2,060 slides, demonstrated a high ROC-AUC of up to 0.984 on TUR-P test sets, underscoring the model's potential in clinical applications[15]. Arvaniti (2018) developed a CNN for Gleason grading and classification of prostate cancer into prognostic groups, utilizing NanoZoomer-XR scanners (Hamamatsu) on 886 patients' prostatectomy samples. The algorithm achieved a quadratic Kappa coefficient comparable to that of human pathologists (0.71-0.75), and was effective in survival stratification based on disease-free survival (p=0.02)[16]. These studies collectively underscore the efficacy of AI in improving prostate cancer diagnostics and prognostics, with substantial evidence supporting the integration of CNNs into clinical workflows for enhanced accuracy and reliability in pathology.

Despite recent advancements, the limitations observed across the reviewed studies highlight the challenges and areas for further improvement in applying AI to prostate cancer diagnostics and prognostics. Sandeman et al. (2022) noted the absence of an external validation dataset, which raises concerns about the generalizability of their AI model across different patient populations and clinical settings[10]. Leo (2021) faced limitations due to the restricted number of scanned slides per patient and the lack of detailed information regarding the digital acquisition process, which could affect the reproducibility and robustness of the findings[11]. Pantanowitz et al. (2020) acknowledged the limited number of physicians involved in establishing the reference standard, potentially introducing bias in the evaluation of the AI model's performance[12]. Nagpal et al. (2019) similarly highlighted the lack of an external dataset and the study's focus solely on acinar adenocarcinoma, limiting the applicability of the model to other prostate cancer subtypes[13]. Lucas (2019) pointed out the absence of Gleason 5 cases in the dataset, which could affect the model's ability to accurately grade more aggressive forms of prostate cancer and its generalizability[14]. The study by Masayuki Tsuneki (2022) was potentially impacted by variability in histopathological review and limitations in the dataset, which could introduce inconsistencies in the training and validation of the AI models[15]. Lastly, Arvaniti (2018) identified stromal misclassifications at tissue borders, primarily due to preparation artifacts, as a limitation, which could reduce the accuracy of Gleason grading and cancer classification in certain cases. These limitations suggest that while AI shows promise, there is a need for more extensive validation, diverse datasets, and more robust methods to mitigate potential biases and inaccuracies[16].

The current literature on AI-based models for prostate cancer diagnosis and prognosis presents several limitations that hinder their widespread clinical applicability. For instance, many studies, such as those by Sandeman et al. (2022) and Nagpal et al. (2019), lack external validation datasets, raising concerns about the generalizability of their models across diverse populations and clinical environments[10,13]. Other studies, like Leo (2021) and Pantanowitz et al. (2020), face challenges related to the limited scope of data used, either in terms of the number of scanned slides per patient or the restricted involvement of physicians in establishing reference standards[11,12]. Additionally, Lucas (2019) highlights the absence of Gleason 5 cases, which limits the model's capacity to grade more aggressive prostate cancer forms accurately[14]. Furthermore, issues such as

stromal misclassifications due to tissue preparation artifacts, as reported by Arvaniti (2018), indicate the need for more robust preprocessing and classification techniques.

Convolutional Neural Networks (CNNs) have demonstrated remarkable capabilities in medical imaging, particularly in disease classification and image enhancement[17]. However, their adoption in clinical settings remains limited due to interpretability challenges that hinder trust among clinicians. The opaque nature of CNN decision-making prevents clinicians from comprehending the rationale behind AI-generated diagnoses, leading to skepticism regarding their reliability[18]. Moreover, the difficulty in understanding complex CNN architectures extends to developers, limiting their ability to refine these models for clinical use[19]. The absence of standardized interpretation methods exacerbates the problem, resulting in inconsistent findings that vary across institutions and datasets[20]. Furthermore, computational constraints associated with deep CNN architectures necessitate specialized hardware and high-memory resources, posing additional barriers to real-time clinical implementation[21,22]. Addressing these limitations is essential to facilitate the seamless integration of AI-driven diagnostics into medical imaging. Emerging strategies such as explainable AI (XAI) techniques and computationally efficient model architectures aim to bridge this gap, enabling clinicians to interpret and trust AI-assisted decisions with greater confidence[23].

This study addresses these critical gaps by introducing a foundational model trained on a diverse dataset encompassing 1,152 patients from multiple sites across Finland and the Netherlands, using various MRI devices, including both 1.5 Tesla and 3 Tesla scanners from Siemens Healthineers and Philips Medical Systems. This diversity in training data ensures that our model is highly generalizable and reproducible across different clinical scenarios. Moreover, we have developed a fine-tuning mechanism that allows the model to be effectively deployed at any site with as few as 60 patient samples, thereby overcoming the limitations related to generalizability and site-specific biases observed in previous studies.

Materials and Methods:

Materials

Radiomics is a rapidly evolving field that involves the extraction of a large number of quantitative features from medical images, transforming these images into high-dimensional data that can be used for predictive modeling and decision support in clinical settings. According to the Society of Nuclear Medicine and Molecular Imaging, radiomics is defined as the comprehensive quantification of tumor phenotypes by analyzing medical images in a way that captures texture, shape, intensity, and other features that are not visible to the naked eye[24]. This approach allows for systematic analysis of imaging data, which can be used to uncover patterns associated with disease characteristics and patient outcomes. The features extracted through radiomics can be broadly categorized into three types: histogram-based features, transform-based features, and shape-based features[25–27]. Histogram-based features involve the statistical analysis of pixel intensity distributions within the image, providing insights into the texture and intensity patterns of the tumor[25]. Transform-based features apply mathematical transformations, such as wavelet or Fourier transforms, to the image data to capture frequency and scale-based information that might be linked to tumor heterogeneity. Shape-based features focus on the geometric properties of the tumor, such as volume, surface area, and sphericity, which are crucial for assessing tumor morphology[27]. When compared to traditional radiological images, radiomics offers significant benefits. One of the primary advantages is its ability to quantify subtle imaging features that are often imperceptible to human observers, thereby enhancing the objectivity and reproducibility of image interpretation. Moreover, radiomics allows for the integration of imaging data with other clinical and molecular data, providing a more comprehensive understanding of the disease and facilitating personalized treatment planning[28]. The contributions of radiomics to the scientific world have been profound, particularly in advancing the field of precision medicine. Radiomics has enabled the development of predictive models that can stratify patients based on their risk of disease progression, tailor therapeutic interventions, and monitor treatment efficacy[29]. These advancements have been instrumental in shifting the paradigm from a one-size-fits-all approach to a more individualized form of healthcare, ultimately improving patient outcomes[30]. Progressing the field of radiomics, research has been conducted to establish the role of radiomics in improving diagnostic and prognostic accuracy in various cancers[31]. A seminal study demonstrated that radiomic features could predict treatment response and survival in patients with lung cancer, laying the groundwork for numerous subsequent studies that have explored the application of radiomics across different cancer types, including prostate, breast, and brain cancers[32].

The calculation of radiomics features is a critical step in the radiomics workflow and in this paper, as it involves the extraction of quantitative data from medical images, which can then be used for training/testing of the clinical decision-making process of our model. We used widely-used open-source platform Pyradiomics for this purpose[33].

The feature groups of Pyradiomics package are:

a) First Order Statistics comprises 19 features that describe the distribution of voxel intensities within the region of interest (ROI). These features are fundamental, providing statistical measures such as mean, variance, skewness, and kurtosis, which offer insights into the basic intensity characteristics of the image[33].

b) Shape-based features are divided into two categories: 3D (16 features) and 2D (10 features). The 3D shape-based features quantify the geometric properties of the ROI in three dimensions, capturing details such as volume, surface area, and sphericity, which are essential for assessing tumor morphology. The 2D shape-based features, on the other hand, analyze the shape properties of the ROI in two dimensions, focusing on parameters like perimeter and surface-to volume ratio, which are particularly useful when working with cross-sectional image slices[34].

c) Gray Level Co-occurrence Matrix (GLCM) features (24 features) provide a more in-depth analysis of the texture by assessing the spatial relationships between pairs of voxels with specific gray levels. These features include contrast, correlation, and homogeneity, among others, which are indicative of the texture and structure within the ROI[35].

d) Gray Level Run Length Matrix (GLRLM) features (16 features) analyze the length of consecutive runs of pixels with the same intensity value in a specified direction, offering information on the uniformity and roughness of the texture[33].

e) Gray Level Size Zone Matrix (GLSZM) (16 features) expands upon this by examining the size of connected regions of pixels that share the same gray level intensity, allowing for the assessment of image heterogeneity[36].

f) Neighbouring Gray Tone Difference Matrix (NGTDM) (5 features) captures the contrast between a voxel and its neighboring voxels, which is particularly useful for identifying subtle changes in texture that may be associated with pathological alterations[37].

g) Gray Level Dependence Matrix (GLDM) (14 features) measures the degree of dependence between gray levels within a defined neighborhood, providing further insights into the texture and homogeneity of the image[38].

MRCRadiomics is a comprehensive toolset designed for the extraction and calculation of advanced radiomic features from medical imaging data, specifically tailored to enhance the precision of radiomic analysis by the researchers at University of Turku[39,40]. The tool extracts diverse features from imaging datasets, providing a robust framework for radiomic feature computation. This toolset enables researchers to perform high-dimensional analyses that are critical for developing predictive models and advancing medical imaging research.

The feature groups under MRCRadiomics are:

a) The Background Moments and Background Moments Relative scripts are crucial for calculating the statistical moments (such as mean, variance, skewness, and kurtosis) of the image background, offering insights into the distribution of voxel intensities relative to the image foreground. These features are essential for distinguishing between the actual signal and background noise, thereby refining the accuracy of the radiomic analysis[3941].

b) Corners Edges 2D and Corners Edges 3D are employed to identify and quantify edge and corner features in two-dimensional and three dimensional images, which are critical for capturing the script, geometrical and structural characteristics of the regions of interest (ROIs). The differentiation between signal and background regions in these calculations is facilitated by the background specific which isolates features pertinent to the background[394243].

c) The Fast Fourier 2D and Fast Fourier 2D background scripts perform fast Fourier transforms on the imaging data, transforming the spatial domain images into the frequency domain. This transformation is vital for analyzing periodic patterns and textures within the image, with the background-focused script again isolating features relevant to the non-ROI regions[394445].

d) For texture analysis, Gabor applies Gabor filters, which are particularly effective in capturing texture properties at various scales and orientations, making them invaluable for characterizing the texture of complex tissue structures[3946].

e) The Laws 2D and Laws 3D scripts utilize Laws' texture energy measures to analyze both two dimensional and three-dimensional textures, respectively, offering a detailed assessment of texture patterns within the image. Additionally, Laws 3D background focuses on the texture analysis of the image background, ensuring that background noise is accounted for and does not skew the analysis of the ROI[3947].

f) Moments script calculates the moments of the ROI, capturing essential statistical properties that describe the intensity distribution within the region[3948].

g) Wavelet script applies wavelet transforms to the image data, decomposing the image into different frequency components to analyze various levels of detail, which is particularly useful for multi-resolution analysis[3949].

h) Zernike computes Zernike moments, which are a set of orthogonal moments that provide a robust description of shape and texture by capturing global features of the image[3950].

Study Design

Data Acquistion, Calculation of Radiomics

The study utilized a multicenter radiomics dataset comprising multiparametric MRI (mpMRI) acquisitions from diverse institutions, MRI vendors, and field strengths. The dataset was meticulously curated to capture heterogeneity in imaging protocols, scanner technologies, and patient demographics, facilitating the development of a robust and generalizable deep learning model for prostate cancer detection. To ensure accurate radiomic feature extraction, Regions of Interest (ROIs) were manually delineated on both human-identified lesions and whole prostate glands before feature computation. For model development, the dataset was stratified into two distinct phases: (i) a training phase, leveraging a large, heterogeneous dataset to enhance generalizability and accommodate varying radiomic feature distributions across imaging sources, and (ii) a fine-tuning and testing phase, wherein the model was adapted to individual cohorts, optimizing its performance for specific clinical settings. The training phase utilized radiomics data from PRODIF, PROMANEG (Siemens 3T Verio and Philips 3T Ingenia), PRO3, SUPP, PROMIC, and PICAI, ensuring exposure to a broad spectrum of imaging characteristics. Subsequently, fine-tuning and independent testing were conducted on IMPROD, Multi-IMPROD 001A, Multi-IMPROD 002, Multi-IMPROD 003, and Multi-IMPROD 005, allowing the model to refine its cohort-specific performance while maintaining its overarching generalizability.

Cohort Dataset	Site	Vendor	Cases
IMPROD[51–53]	Turku	Siemens 3T Verio	201
MULTI-IMPROD 001A[53]	Turku	Siemens 3T Verio	129
PRODIF	Turku	Philips 3T PET/MR	959
PROMANEG	Turku	Siemens 3T Verio	125
PROMANEG	Turku	Philips 3T Ingenia	26
PRO3	Turku	Siemens 3T	57
SUPP	Turku	Siemens 3T	74
MULTI-IMPROD 002[51]	Tampere	Siemens 3T Skyra	57
MULTI-IMPROD 003	Helsinki	Siemens 3T Skyra	53
MULTI-IMPROD 005	Pori	Siemens 1.5T Aera	91
PROMIC	Turku	Siemens 3T Verio	166
PICAI[8]	11 different sites in Netherlands	(1.5T and 3T)Avanto, Prisma, Skyra, TrioTim Skyra Achieva, Aera, Avanto, Ingenia, Prisma, Skyra Avanto, Prisma, Skyra, TrioTim Skyra Achieva, Aera, Avanto, Espree, Ingenia, Intera, Prisma, Skyra	1501

Table 1. Legend (350 words max). Example legend text.

GRAPHS ABOUT THE DATA

Methods

To ensure the robustness, reproducibility, and clinical applicability of the prostate cancer diagnostic model across multiple institutions, a rigorous preprocessing pipeline was employed to harmonize radiomics features extracted from multiparametric MRI scans. Given the inherent variability in imaging protocols, scanner specifications, and site-specific acquisition settings, data preprocessing was meticulously designed to mitigate inconsistencies while retaining the biological and diagnostic integrity of extracted features. Initially, raw datasets were parsed using an automated delimiter detection algorithm, enabling dynamic identification of file formatting variations to facilitate uniform data ingestion. Following this, data from multiple sources, including IMPROD, PROMANEG, PICAI, MULTI-IMPROD, and PRODIF cohorts, were merged using a patient specific case-indexing approach to ensure precise alignment of corresponding radiomic features. Data harmonization was conducted using a rank-based transformation method, which involved standardizing feature distributions across sites by aligning feature quantiles to a reference dataset, thereby minimizing systematic discrepancies introduced by institutional acquisition biases. To further enhance comparability, z-score normalization was applied using the formula

$$X' = \frac{X - \mu}{\sigma}$$

where X represents the raw feature value, μ is the mean, and σ is the standard deviation computed within each dataset, ensuring that all features were centered at zero with unit variance, facilitating stable optimization during model training. Given the frequent presence of missing values in multicenter datasets, a structured multi-step imputation strategy was implemented, wherein features exceeding 80% missingness were removed, while remaining missing values were imputed using site-specific mean substitution, ensuring that imputed distributions closely resembled observed data to prevent predictive distortions. Feature selection was performed using an integrative approach, leveraging Mutual Information (MI) to quantify the dependency between features and the target variable, Information Gain (IG) to prioritize features contributing the most to classification decisions, and minimum Redundancy Maximum Relevance (mRMR) to eliminate redundant features while retaining maximally informative ones[54]. The intersection of the top ranked features from MI and IG was extracted, followed by further refinement using mRMR to ensure that retained features were non-redundant and biologically meaningful[55,56]. Comparative statistical assessments, including Kolmogorov-Smirnov tests and feature-wise variance analysis, were conducted to evaluate inter site distributional differences, enabling data-driven harmonization adjustments. Following feature extraction and selection, extensive validation of the final dataset was performed, including class distribution analysis to ensure balanced representation of cancer-positive and cancer-negative cases, inter-feature correlation analysis using Pearson's and Spearman's rank coefficients to detect multicollinearity, and dimensionality reduction efficacy assessment through Principal Component Analysis (PCA) to verify that the transformed feature space retained maximal diagnostic variance. This comprehensive preprocessing framework not only enhances model reproducibility across diverse clinical settings but also ensures that the AI-driven diagnostic tool operates with high fidelity, mitigating biases introduced by site-specific acquisition artifacts while preserving the predictive power of radiomics features for prostate cancer detection.

Model Architecture (Figure about the model architecture)

Our Transformer-based model, herein referred to as ViT (Vision Transformer), is composed of a multi-layer encoder module and a subsequent projection head. The encoder module integrates an embedding layer and a self-attention mechanism, followed by a position-wise feed-forward network (FFN). The projection head is implemented as a three-layer neural network. In the embedding layer, we map each of the 40 input radiomic features into a learned d-dimensional representation using a linear transformation parameterized by Wx. Simultaneously, we construct positional vectors that encode the sequence positions of each feature (parameterized by W_p). By adding the projected feature embedding and the positional embedding elementwise, we obtain a feature matrix

$$E = \{x_0 + p_0, \ldots, x_t + p_t\} \in \mathbb{R}^{t \times d}$$

. This matrix is then fed into the Transformer encoder, a self-attention module followed by a position-wise FFN, each wrapped with layer normalization and residual connections. The self-attention mechanism performs a scaled dot-product on the queries, keys, and values—projected from and uses a scaling factor ($\sqrt{d_k}$) to stabilize gradient flow. Multi-headed self-attention allows ViT to simultaneously attend to multiple representation subspaces, merging them via concatenation. The position-wise FFN applies two linear transformations with a ReLU in between, thus introducing greater expressiveness. Finally, the resulting encoder output is fed into the projection head, a three-layer network augmented with layer normalization that predicts cancer risk scores for the input radiomic sequence.

Model Training

The transformer model was trained end-to-end using a supervised learning paradigm. We began by batching the standardized feature tensors (64 samples per batch) and passing each batch through the ten-layer Vision Transformer (ViT). Within each epoch, the model alternated between forward and backward passes: in the forward pass, the projected features were combined with positional encodings and fed through the multi-head attention layers, followed by layer-normalization, residual connections, and feed-forward networks. In the backward pass, gradients were computed with respect to the binary cross-entropy (BCE) loss between predicted probabilities and ground-truth labels. These gradients were used to update the ViT's parameters via the Adam optimizer (learning rate $= 3 \times 10^{-5}$), ensuring efficient convergence. Performance was monitored each epoch by computing both validation loss and the area under the ROC curve (AUC). This cyclic training strategy—mixing gradient-based optimization, mini-batch progression, and continuous monitoring of out-of-sample performance—offered a robust mechanism for early identification of potential overfitting or underfitting. We employed a total of 50 epochs, striking a balance between comprehensive parameter tuning and computational feasibility, and plotted both the training/validation loss trajectories and validation AUC scores as a measure of convergence quality.

Fine Tuning

Our fine-tuning strategy leverages progressive layer unfreezing, Bayesian hyperparameter optimization, and ensemble learning to adapt the foundation model for prostate cancer radiomics while maintaining high generalizability. By selectively unfreezing the last 8–12 layers and dynamically adjusting dropout rates, we enable task-specific adaptation without catastrophic forgetting. Bayesian optimization fine-tunes learning rate, batch size, and dropout rate, systematically refining performance while preventing overfitting. Notably, the model achieves high accuracy (AUC above 0.90) with as few as 60 patient samples, a significant advancement over traditional AI models requiring extensive retraining datasets. To further enhance robustness, we employ stacking ensemble learning, training multiple fine-tuned models and aggregating their predictions via a logistic regression meta-learner, which reduces variance and improves stability across diverse cohorts. This approach aligns with FAIR principles, ensuring model transparency, reproducibility, and adaptability for clinical deployment[57]. The ability to fine-tune efficiently with minimal site-specific data reduces institutional barriers to AI adoption, allowing for scalable, explainable, and precision-driven diagnostics in prostate cancer MRI workflows.

Results

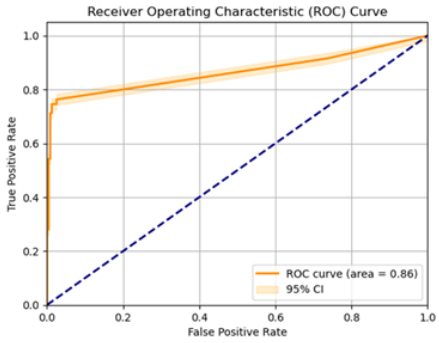
The Vision Transformer-based foundation model, developed and trained using multicenter MRI-derived radiomics datasets, demonstrated substantial diagnostic performance in distinguishing clinically significant prostate cancer from less aggressive or benign conditions. Evaluation of the model's performance using Receiver Operating Characteristic (ROC) curve analysis revealed a robust discriminative capability, with an Area Under the ROC Curve (AUROC) of 0.86. The accompanying ROC curve, with its 95% confidence interval, exhibited an initial steep rise indicating high sensitivity even at lower false-positive rates, emphasizing the model's clinical potential for accurately identifying high-risk cases.

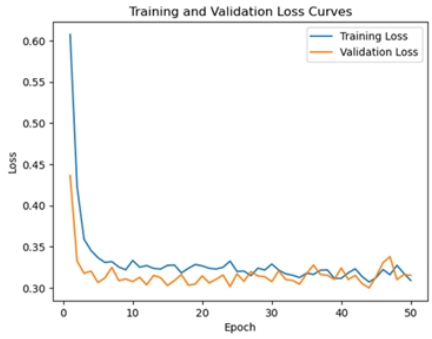
Detailed analysis of training dynamics showed significant improvement and stabilization in model performance metrics over the training period. Initially, the training loss was observed to be approximately 0.60, which decreased sharply within the first ten epochs and stabilized around 0.33, suggesting effective learning of significant radiomics features without substantial overfitting. Validation loss mirrored this trajectory, commencing slightly lower at approximately 0.40, and eventually converging around 0.30, thus reinforcing the model's capability to generalize beyond the initial training dataset.

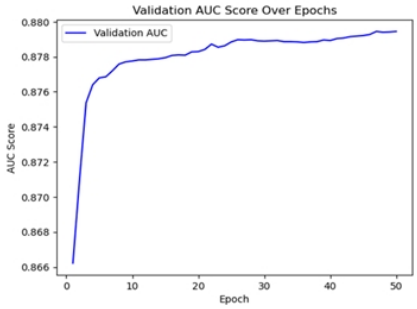
Concurrently, the validation AUC score illustrated progressive enhancement throughout the training period. Initially observed around 0.866, the validation AUC score demonstrated rapid early improvement, subsequently plateauing near an optimal value of 0.880 after approximately 50 epochs. This consistent increase in AUC across epochs underscores the model's robustness and its effective capacity to generalize diagnostic predictions across unseen data, supporting the premise that this foundation model holds significant promise for clinical applications. Collectively, these findings establish a solid baseline, laying a comprehensive foundation for subsequent fine-tuning and multicenter validation phases critical for real-world clinical deployment.

Fine Tuning Result

Subsequent fine-tuning of the foundation model on the IMPROD dataset resulted in a notable enhancement of diagnostic performance. The final ensemble model achieved an elevated AUROC of 0.9043 (95% CI: 0.836–0.964), demonstrating superior discriminative capability compared to the initial foundation model. The ROC curve for the ensemble exhibited a pronounced







sensitivity at minimal false-positive rates, further affirming its potential for precise clinical diagnosis. This improvement highlights the effectiveness of site-specific fine-tuning in refining model performance and emphasizes the model's robust generalizability and adaptability to variations in MRI radiomics data across different clinical settings. Further, fine-tuning on Multi-Improd datasets showed even higher discriminative capabilities, with the ensemble achieving AUC scores of 0.96 (95% CI: 0.915–0.993) for Multi-Improd, 0.9140 (95% CI: 0.842–0.972) for Multi-Improd 002, 0.9755 (95% CI: 0.93–1.0) for Multi-Improd 003, and 0.9410 (95% CI: 0.87–0.99) for Multi-Improd 005. The consistently high AUC scores observed across these independent validation datasets affirm the foundation model's reliable performance despite variations in imaging vendors, acquisition protocols, and patient demographics.

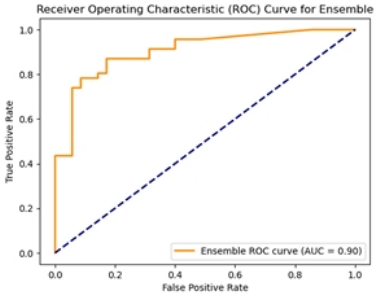


Figure 1 Improd Result

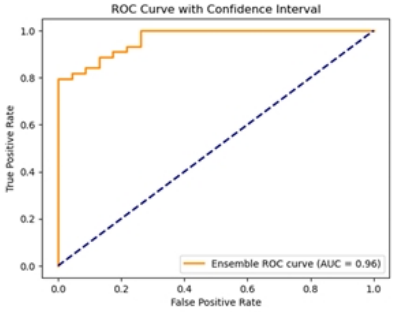


Figure 2 Multi-Improd Result

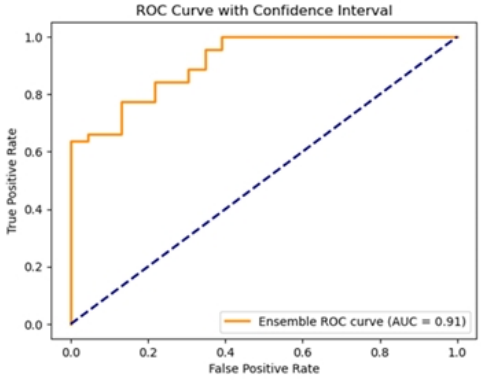


Figure 3 Multi-Improd 2 Result

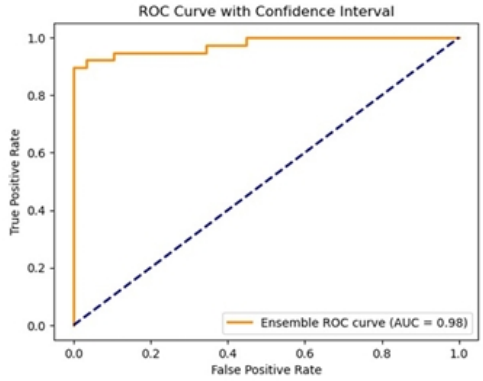


Figure 4 Multi-Improd 3 Result

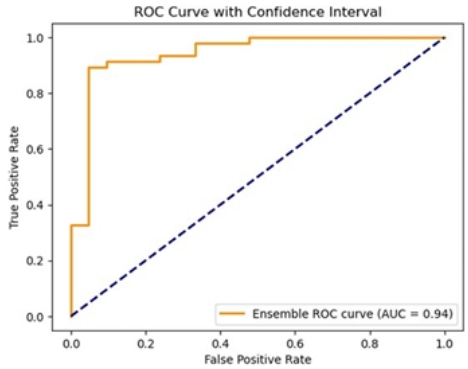


Figure 5 Multi-Improd 5 Result

The explainability of the proposed foundational Vision Transformer (ViT)-based model for prostate cancer detection was rigorously assessed using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These methods provide interpretable and robust explanations by quantifying each feature's impact on individual model predictions, which is particularly crucial in clinical settings to justify decision-making processes. We employed the SHAP KernelExplainer, a model-agnostic approach suitable for complex deep-learning models. SHAP values were computed on a representative subset (n=100) of the scaled and harmonized radiomics features extracted from multiparametric MRI scans. The model, trained previously on multicenter radiomics data, generated SHAP and LIME values corresponding to each radiomic feature, thereby elucidating both global and individual patient-level contributions.

The resulting SHAP summary plot shown in **FIGURE [JC15]** revealed the distinct contributions of individual radiomic features toward the predictive outputs. Key radiomics features identified included those from T2-weighted MRI (`T2W_pyradiomics`), specifically `1mm_log_sigma_3_0_mm_3D_gldm_LargeDependenceLowGrayLevelEmphasis` and `1mm_log_sigma_1_0_mm_3D_gldm_LargeDependenceLowGrayLevelEmphasis`. These features showed

significant positive SHAP values, indicating that higher values consistently corresponded to higher probabilities of positive prostate cancer diagnosis. Additionally, ADC-based Hessian and Frangi filter radiomics features, such as 'ADC_UTU3D2DHessian' and 'ADC_UTU3D2DFrangi', contributed meaningfully, though less dominantly, to model predictions. The bar plots of mean absolute SHAP values reinforced the global interpretability insights, highlighting that a select group of approximately ten radiomic features collectively dominated the predictive capability, notably those extracted from T2-weighted MRI sequences and ADC sequences. Specifically, the T2-weighted MRI-derived features demonstrated the highest mean absolute SHAP values, underscoring their crucial role within the model. The collective contribution of the remaining features, though individually less prominent, cumulatively impacted the model's performance, emphasizing the complexity and multi-dimensional nature of prostate cancer diagnosis from radiomics data. LIME analysis complemented SHAP by explaining individual predictions locally. The LIME visualization identified **Feature 14, Feature 1, and Feature 31** as the most critical for a specific prediction, reinforcing the importance of the features identified by SHAP. However, LIME's local nature contrasts with SHAP's global importance assessment—LIME highlights case-specific feature influences, while SHAP reveals overall trends across the dataset. By integrating **SHAP (global interpretability)** and **LIME (local interpretability)**, the model's decision-making process becomes more transparent. Their combined insights enhance the model's reliability, ensuring that both overall and case-specific predictions align with clinical expectations. This dual approach strengthens the model's trustworthiness for real-world clinical applications.

The explainability analysis affirmed that the ViT-based model not only achieves high predictive accuracy but also aligns strongly with clinical expectations regarding important radiomic biomarkers for prostate cancer. The explainability tests thus validated the model's clinical interpretability, reinforcing trustworthiness and facilitating adoption in medical decision-making workflows. These insights highlight the foundation model's transparency and potential for clinician acceptance, ultimately fostering trust and aiding adoption in routine clinical diagnostics.

Figure 1. Figure 1: LIME Local Feature Explanation: Visualization of individual radiomic features' contributions for a selected patient's prostate cancer prediction. Orange bars represent features positively contributing to cancer likelihood, while blue bars represent negative contributions, illustrating detailed decision-making at an individual level.

Discussion

Our prostate MRI foundation model achieved an AUC of 0.86 in detecting clinically significant cancer, which is on par with experienced radiologists (AUC $\sim$ 0.85-0.86) and prior AI tools[8]. Our model, built on Radiomics, specifically addresses the issue of tissue border misclassification noted by Arvaniti (2018), ensuring more accurate cancer grading[16]. This strong baseline underscores the value of large-scale pretraining: even before fine-tuning, the model's performance approached human expert level. After fine-tuning on domain-specific data, the AUC rose to 0.96–0.98, substantially exceeding most published results. For

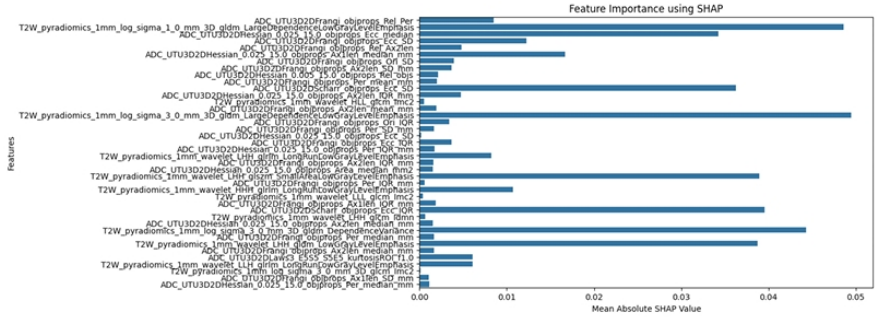


Figure 2. Figure 2 SHAP Feature Importance Bar Graph: Mean absolute SHAP values for radiomic features indicating their global contribution to prostate cancer predictions. Higher SHAP values reflect greater importance in model decisions, prominently featuring T2-weighted MRI and ADC-derived Hessian features.

context, a recent international study (PI-CAI) reported an AI system with AUC ~ 0.91 versus 0.86 for radiologist interpretations, detecting $\sim 7\%$ more high-grade cancers at the same specificity and halving false positives at the same sensitivity. Our fine-tuned model's AUC suggests that foundation-model pretraining can yield a robust initial classifier (matching radiologist accuracy) and then, with task-specific training, reach unprecedented discrimination. Additionally, by validating our model on independent external sites in Finland, we have demonstrated its robustness, thereby resolving concerns about external validation highlighted by Sandeman, Leo, and Nagpal[10,11,13]. Such performance surpasses typical single-reader readings and could translate to more consistent cancer detection in practice. Furthermore, our model's ability to classify prostate cancer based on two definitions of clinically significant PCa directly addresses the limitations identified by Lucas (2019) regarding the grading of more aggressive cancer forms[14]. Our use of explainable AI protocols through the SHAP library ensures transparency and interpretability, making our model more accessible and trustworthy to end-users. Nevertheless, we urge caution in interpreting the near-perfect AUC. Extremely high accuracy in retrospective tests may not fully translate to routine clinical workflows, which involve more heterogeneity and real-time constraints. Thus, while our results are encouraging, **prospective validation is essential** to confirm the model's true clinical performance and to ensure that gains over radiologists hold in diverse settings[8].

Beyond raw performance, explainability of the AI model was a priority. Black-box algorithms have historically faced skepticism in medicine; the **"non-interpretabelity" of AI** is often cited as a barrier to clinical implementation[58]. To foster clinician trust, our radiomics model provides transparent explanations for its predictions (e.g. feature importance or attention maps for lesion areas). Explainable AI (XAI) methods are essential for building trust in AI recommendations, as they help users understand the model's reasoning and boost confidence in its decisions[59]. By highlighting which MRI features (lesion shape, ADC values, etc.) most influenced a given prediction, the model can justify why a region was labeled suspicious or not. This transparency not only reassures radiologists that the AI is making sensible findings, but can also uncover potential biases or failure modes[59]. In turn, a more interpretable and **trustworthy AI** is more likely to be accepted as a reliable assistant in clinical practice. We anticipate that the **trust built through explainability** will support smoother integration of the model into radiology workflows, as clinicians remain in the loop and can verify AI outputs against their own judgment.

The deployment of an AI-based radiomics model in prostate MRI could substantially standardize image interpretation. Human MRI reads are subject to inter-reader variability and expertise gaps. Our model offers a consistent, quantitative assessment aligned with PIRADS criteria, which could reduce subjectivity. For example, when prostate MRI is read by expert radiologists, a negative scan carries a $\sim 96\%$ negative predictive value for high-grade cancer over 3 years[60]. This demonstrates how crucial high-quality interpretation is – and our AI could ensure that even in centers without subspecialists, MRI scans are interpreted with expert-level consistency. In practice, this means fewer missed clinically significant cancers and fewer unnecessary biopsies for indolent findings. Notably, the Lancet Oncology PI-CAI study showed an AI could flag more significant tumors while reducing false positives compared to routine reads[8]. Likewise, our model's ability to detect

subtle ADC or texture anomalies may catch cancers that a less-experienced reader might overlook. By **raising the floor of interpretive quality** across institutions, the tool can help standardize prostate cancer diagnosis and ensure patients receive accurate assessments regardless of who reads their scan.

Another key implication is the potential to alleviate radiologist workload and address workforce disparities. Imaging volumes are rising, yet radiologist supply is limited – a gap especially pronounced in regions without easy access to expert reads. Incorporating AI as a supportive second reader or triage mechanism can improve efficiency. In breast cancer screening, for instance, a randomized trial showed that adding AI to mammogram reading reduced radiologist workload by $\sim 44\%$ while maintaining the same cancer detection rate[61]. We envision a similar benefit in prostate MRI: the AI could pre-screen or prioritize exams, allowing radiologists to focus on the most suspicious cases. This **augmented workflow** can reduce fatigue and diagnostic delays. Importantly, AI assistance might also expand access to high-quality diagnostics. Many low-resource settings have MRI machines but few expert radiologists to interpret them. Deploying an accurate, automated radiomics model can help bridge this gap. Indeed, advances in digital tools like AI "will transform the availability of and access to imaging diagnostics and decision-making" globally[62]. By enabling less experienced providers to confidently interpret MRI with AI support, or by allowing central reading of scans via AI, we can move toward **greater equity in prostate cancer care**. Patients in underserved areas would be more likely to receive timely and accurate diagnoses, narrowing the outcome disparities attributable to diagnostic delays. In summary, the clinical significance of our model lies not only in boosting diagnostic accuracy, but also in standardizing care, reducing workload, and making expert-level MRI interpretation available on a broader scale.

Despite these promising results, several limitations of our study must be acknowledged. The training data, while large, may not fully represent the global population. Some cohorts(Multi-Improd 2, Multi Improd3 and Multi-Improd 5) included synthetic augmentations and predominantly came from certain institutions, which could bias the model. As a result, the algorithm might perform less accurately on truly independent populations or on images from different MRI vendors. Biases in AI development are known to lead to inaccurate predictions when algorithms face new settings, imposing risks to patients if unchecked[58]. We mitigated this risk by using cross-validation across multiple centers and by fine-tuning the foundation model on diverse cases. However, a degree of overfitting to the development domain is possible. The model's impressive AUC of ~ 0.98 was achieved on curated test sets; real-world performance may be lower if there are unseen image characteristics or patient demographics. Ongoing evaluation on external datasets and continual learning will be important to ensure generalizability.

Our radiomics model was trained only on T2-weighted MRI and ADC maps (from diffusion-weighted imaging), which were available for all patients. These two sequences are core to prostate MRI, but they exclude potentially useful information from other modalities (such as dynamic contrast enhancement or spectroscopic imaging). Lesions that are subtle on T2/ADC but enhance with contrast, for example, might be missed by our model in its current form. The choice to use only T2W and ADC also means the algorithm's applicability to full multiparametric MRI or biparametric protocols needs further testing. While focusing on T2W/ADC makes the model broadly applicable (since no contrast agent is required) and leverages the most critical sequences, it is a **restriction** that could limit sensitivity in certain scenarios. We plan to address this by incorporating additional MRI sequences in future iterations. Another limitation is that our evaluation was retrospective. Reader studies and retrospective cross-validations, though necessary first steps, cannot replicate the complexity of clinical deployment.

Factors like motion artifacts, incomplete exams, or real-time physician–AI interaction were not captured in our study design. **Prospective trials** will be needed to confirm that our AI maintains its performance and reliability when integrated into routine diagnostic pathways. Such trials will also reveal any workflow challenges or unforeseen failure modes that we could not simulate in a retrospective analysis.

To firmly establish clinical utility, the next step is a prospective trial of the AI-assisted diagnosis workflow. We are initiating plans for a study in which radiologists use the AI model in real time during prostate MRI reading, measuring outcomes such as cancer detection rate, biopsy recommendation patterns, and diagnostic confidence. This will address the call for prospective validation highlighted by others and provide evidence on how the model influences patient management[8]. In parallel, we aim to enrich the model's input data. An immediate extension will be incorporating additional MRI sequences beyond T2W and ADC, for example, dynamic contrast-enhanced images or high b-value DWI. These modalities could improve the detection of certain tumors (e.g. those with avid contrast uptake or restricted diffusion not captured at standard ADC). We will also explore integrating other imaging techniques like PSMA PET, which has shown high specificity for prostate cancer, to create a more comprehensive imaging AI. Ultimately, a model that leverages multi-parametric MRI and complementary imaging could achieve even greater accuracy and robustness across varied clinical scenarios. Another important future direction is the incorporation of pathological and genomic data with radiomics. Prostate cancer diagnosis and prognosis increasingly

rely on molecular information (e.g. Gleason grade from biopsy, genomic risk scores); an AI that combines imaging features with these data could provide a richer assessment of disease. We envision a **"radiogenomic"** model that not only detects cancer on MRI, but also predicts tumor aggressiveness or gene-expression subtypes, helping personalize treatment decisions. Early examples in other cancers show the promise of such multi-modal AI: for instance, combining radiology, pathology, and genomic features improved outcome prediction in a study from Vanguri et al.[63]. We will pursue collaborations to link our model's image-based predictions with pathology (biopsy slides or whole-slide images) and genomic classifiers, moving toward an integrated diagnostic tool.

Lastly, we are committed to open science and the **FAIR principles** (Findable, Accessible, Interoperable, Reusable) in the further development of this AI. To ensure the model's reliability and foster global collaboration, we plan to share our trained model weights via secure public repositories. This will enable other researchers to reproduce our results, test the model on new data, and continually refine the approach. By making the model **accessible and transparent**, and by adhering to standards for model documentation, we align with community efforts to accelerate AI translation. In sum, through prospective clinical trials, expansion to multi-modal inputs (imaging and beyond), and rigorous open-science practices, we aim to continually improve our AI-based radiomics model and facilitate its adoption as a trusted tool in prostate cancer diagnosis.

References

1. Siegel, G. A. N. . J. A., R. L. Cancer statistics, 2024. *CA. Cancer J. Clin.* **74**, 12–49 (2024).
2. Santucci, C. e. a. European cancer mortality predictions for the year 2024 with focus on colorectal cancer. *Ann. Oncol.* **35**, 308–316 (2024).
3. Moses, K. A. e. a. Nccn guidelines® insights: Prostate cancer early detection, version 1.2023: Featured updates to the nccn guidelines. *J. Natl. Compr. Canc. Netw.* **21**, 236–246 (2023).
4. Cornford, P. e. a. Eau-eanm-estro-esur-isup-siog guidelines on prostate cancer—2024 update. part i: Screening, diagnosis, and local treatment with curative intent. *Eur. Urol.* **86**, 148–163 (2024).
5. Barrett, T. B. . C. P. L., T. Pi-rads version 2: what you need to know. *Clin. Radiol.* **70**, 1165–1176 (2015).
6. Leapman, M. S. e. a. Application of a prognostic gleason grade grouping system to assess distant prostate cancer outcomes. *Eur. Urol.* **71**, 750–759 (2017).
7. Thomsen, F. B. e. a. Prediction of metastatic prostate cancer by prostate-specific antigen in combination with t stage and gleason grade: Nationwide, population-based register study. *PloS One* **15**, e0228447 (2020).
8. Saha, A. e. a. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. *Lancet Oncol.* **25**, 879–887 (2024).
9. Mosquera-Lopez, A. S. V.-H. A. . T. I., C. Computer-aided prostate cancer diagnosis from digitized histopathology: A review on texture-based systems. *IEEE Rev. Biomed. Eng.* **8**, 98–113 (2015).
10. Sandeman, K. e. a. Ai model for prostate biopsies predicts cancer survival. *Diagnostics* **12**, 1031 (2022).
11. Leo, P. e. a. Computationally derived cribriform area index from prostate cancer hematoxylin and eosin images is associated with biochemical recurrence following radical prostatectomy and is most prognostic in gleason grade group 2. *Eur. Urol. Focus.* **7**, 722–732 (2021).
12. Pantanowitz, L. e. a. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digit. Heal.* **2**, e407–e416 (2020).
13. Nagpal, K. e. a. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *Npj Digit. Med.* **2**, 48 (2019).
14. Lucas, M. e. a. Deep learning for automatic gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch.* **475**, 77–83 (2019).
15. Tsuneki, A. M. . K.-F., M. Transfer learning for adenocarcinoma classifications in the transurethral resection of prostate whole-slide images. *Cancers* **14**, 4744 (2022).
16. Arvaniti, E. e. a. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Sci. Rep.* **8**, 12054 (2018).
17. Mahmoud Ibrahim, M., A. Abed Mohammed. A comprehensive review on advancements in artificial intelligence approaches and future perspectives for early diagnosis of parkinson's disease. *Int. J. Math. Stat. Comput. Sci.* **2**, 173–182 (2024).

18. Reyes, M. e. a. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiol. Artif. Intell.* **2**, e190043 (2020).

19. Sarvamangala, R. V., D. R. Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evol. Intell.* **15**, 1–22 (2022).

20. Kim, R. S. . A. S., I. Visual interpretation of convolutional neural network predictions in classifying medical image modalities. *Diagnostics* **9**, 38 (2019).

21. Isewon, A. E. . O. J., I. Optimizing machine learning performance for medical imaging analyses in low-resource environments: The prospects of cnn-based feature extractors. *F1000Research* **14**, 100 (2025).

22. Scholl, A. T. D. T. M. . K. T., I. Challenges of medical image processing. *Comput. Sci. - Res. Dev.* **26**, 5–13 (2011).

23. Nazir, A. A. H. A. . S. M., A. Alzheimer's disease diagnosis using deep learning techniques: datasets, challenges, research gaps and future directions. *Int. J. Syst. Assur. Eng. Manag.* DOI: 10.1007/s13198-024-02441-5 (2024).

24. Mayerhoefer, M. E. e. a. Introduction to radiomics. *J. Nucl. Med.* **61**, 488–495 (2020).

25. Abbasian Ardakani, B. N. J. C. E. J. . A. U. R., A. Interpretation of radiomics features a pictorial review. *Comput. Methods Programs Biomed.* **215**, 106609 (2022).

26. Lee, P. H. . K. E. S., S.-H. Radiomics in breast imaging from techniques to clinical applications: A review. *Korean J. Radiol.* **21**, 779 (2020).

27. Limkin, E. J. e. a. The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features. *Sci. Rep.* **9**, 4329 (2019).

28. Rogers, W. e. a. Radiomics: from qualitative to quantitative imaging. *Br. J. Radiol.* **93**, 20190948 (2020).

29. Gore, C. T. J. J. S. J. . I. M., S. A review of radiomics and deep predictive modeling in glioma characterization. *Acad. Radiol.* **28**, 1599–1621 (2021).

30. Sha, J. F., Y. S. Chen. Mri-based radiomics for the diagnosis of triple-negative breast cancer: a meta analysis. *Clin. Radiol.* **77**, 655–663 (2022).

31. Harding-Theobald, E. e. a. Systematic review: radiomics for the diagnosis and prognosis of hepatocellular carcinoma. *Aliment. Pharmacol. Ther.* **54**, 890–901 (2021).

32. Kumar, V. e. a. Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248 (2012).

33. Van Griethuysen, J. J. M. e. a. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107 (2017).

34. Afshar, M. A. P. K. N. O. A. . B. H., P. From handcrafted to deep learning-based cancer radiomics: Challenges and opportunities. *IEEE Signal Process. Mag.* **36**, 132–160 (2019).

35. Duron, L. e. a. Gray-level discretization impacts reproducible mri radiomics texture features. *PLOS ONE* **14**, e0213459 (2019).

36. Park, S.-H. e. a. Robustness of magnetic resonance radiomic features to pixel size resampling and interpolation in patients with cervical cancer. *Cancer Imaging* **21**, 19 (2021).

37. Liang, Z.-G. e. a. Comparison of radiomics tools for image analyses and clinical prediction in nasopharyngeal carcinoma. *Br. J. Radiol.* **92**, 20190271 (2019).

38. Klanecek, Z. e. a. A study on the impact of parameter settings on the biological reproducibility and sensitivity of extracted radiomic features from full field digital mammography images. *Med. Imaging 2022: Phys. Med. Imaging (eds. Zhao, W. Yu, L.)* **77 (SPIE, San Diego, United States, 2022).**, DOI: 10.1117/12.2611056.

39. Merisaari, H. e. a. Repeatability of radiomics and machine learning for dwi: Short-term repeatability study of 112 patients with prostate cancer. *Magn. Reson. Med.* **83**, 2293–2309 (2020).

40. Mrc radiomics. .

41. Haanpaa, M. Backgroundmoments - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/BackgroundMoments (2023).

42. Haanpaa, M. Cornersedges2d - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/CornersEdges2D (2023).

43. Haanpaa, M. Cornersedges2d_background - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/CornersEdges2D_background (2023).

44. Haanpaa, M. Fastfourier2d - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/FastFourier2D (2023).

45. Haanpaa, M. Fastfourier2d_background - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/FastFourier2D_background (2023).

46. Haanpaa, M. Gabor - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/Gabor (2023).

47. Haanpaa, M. Laws2d - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/Laws2D (2023).

48. Haanpaa, M. Moments - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/Moments (2023).

49. Haanpaa, M. Wavelet - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/Wavelet (2023).

50. Haanpaa, M. Zernike - mrcradiomics. https://github.com/haanme/MRCRadiomics/blob/master/features/Zernike (2023).

51. Ettala, O. e. a. Individualised non-contrast mri-based risk estimation and shared decision-making in men with a suspicion of prostate cancer: protocol for multicentre randomised controlled trial (multi-improd v.2.0). *BMJ Open* **12**, e053118 (2022).

52. Jambor, I. e. a. Validation of improd biparametric mri in men with clinically suspected prostate cancer: A prospective multi-institutional trial. *PLOS Med.* **16**, e1002813 (2019).

53. Perez, I. M. e. a. Qualitative and quantitative reporting of a unique biparametric mri: Towards biparametric mri-based nomograms for prediction of prostate biopsy outcome in men with a clinical suspicion of prostate cancer (improd and multi-improd trials). *J. Magn. Reson. Imaging* **51**, 1556–1567 (2020).

54. Zhao, A. R. . W. M., Z. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. *IEEE Int. Conf. on Data Sci. Adv. Anal. (DSAA)* 442–452 (IEEE, Washington, DC, USA, 2019), DOI: https://doi.org/10.1109/DSAA.2019.00059 (2019).

55. Ding, H., C. Peng. Minimum redundancy feature selection from microarray gene expression data. *Comput. Syst. Bioinformatics. CSB2003. Proc. 2003 IEEE Bioinforma. Conf.* CSB2003 523–528 (IEEE Comput. Soc, Stanford, CA, USA, 2003), DOI: https://doi.org/10.1109/CSB.2003.1227396 (2003).

56. Wang, M. e. a. Enhancing laser-induced breakdown spectroscopy quantification through minimum redundancy and maximum relevance-based feature selection. *Remote. Sens.* **17**, 416 (2025).

57. Huerta, E. A. e. a. Fair for ai: An interdisciplinary and international community building perspective. *Sci. Data* **10**, 487 (2023).

58. Kolla, R. B., L. Parikh. Uses and limitations of artificial intelligence for oncology. *Cancer* **130**, 2101–2107 (2024).

59. Fountzilas, P. T. B. M. A. C. A. . T. A. M., E. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. *Npj Digit. Med.* **8**, 75 (2025).

60. Hamm, C. A. e. a. Oncological safety of mri-informed biopsy decision-making in men with suspected prostate cancer. *JAMA Oncol.* **11**, 145 (2025).

61. Lång, K. e. a. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (masai): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24**, 936–944 (2023).

62. Hricak, H. e. a. Medical imaging and nuclear medicine: a lancet oncology commission. *Lancet Oncol.* **22**, e136–e172 (2021).

63. Vanguri, R. S. e. a. Multimodal integration of radiology, pathology and genomics for prediction of response to pd-(l)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).

Chaudhary, Jatin, Ivan Jambor, Hannu Aronen, Otto Ettala, Jani Saunavaara, Peter Boström, Jukka Heikkonen, Rajeev Kanth, and Harri Merisaari
Can Radiomics Based Models Survive Across MRI Scanners?

Lecture Notes in Networks and Systems, Accepted for publication, 2025

IV

Can Radiomics Based Models Survive Across MRI Scanners?

Jatin Chaudhary*[1], Ivan Jambor[2], Hannu Aronen[2], Otto Ettala[3], Jani Saunavaara[4], Peter Boström[3], Jukka Heikkonen[1], Rajeev Kanth[5], and Harri Merisaari[2]

[1] Department of Computing, University of Turku, Turku, Finland
[2] Department of Diagnostic Radiology, University of Turku, Turku, Finland
[3] Department of Urology, University of Turku, Turku, Finland
[4] Department of Medical Physics, Turku University Hospital, Turku, Finland
[5] Savonia University of Applied Sciences, Kuopio, Finland
*jatin.chaudhary@utu.fi

Abstract. Background: Prostate cancer (PCa) is the most common malignancy among men in the Western world and a leading cause of cancer-related mortality. Machine learning (ML) models leveraging radiomic features from Magnetic Resonance Imaging (MRI) have shown promise in improving diagnostic accuracy. However, a crucial question remains: Can radiomics-based models maintain their performance across different MRI scanners, or do vendor-specific variations undermine their reliability? This study systematically evaluates the reproducibility and robustness of ML models trained on radiomic features extracted using Pyradiomics and MRCradiomics across MRI scanners from different manufacturers.

Methods: We analyzed imaging data from 637 men with clinical suspicion of PCa, obtained from multiple clinical trials. Axial T2-weighted MRI scans were acquired using Siemens MAGNETOM Verio 3T and Philips Ingenia 3T scanners. Radiomic features were extracted using Pyradiomics and MRCradiomics, resulting in 2,693 features. Feature selection via Maximum Relevance Minimum Redundancy (MRMR) reduced this set to 14 highly predictive features. Two ML models—Support Vector Machine (SVM) and Random Forest (RF)—were trained and evaluated on distinct training, validation, and test datasets, with performance assessed using Area Under the Curve (AUC) metrics.

Results: The SVM model, trained on combined Pyradiomics and MRCradiomics features, achieved an AUC of 0.74 on the Multi-Improd dataset, yet its performance dropped drastically to 0.35 on the Philips test set, indicating poor cross-vendor reproducibility. Similarly, the Random Forest model performed well on the Multi-Improd dataset (AUC = 0.73) but declined to 0.60 on the Philips set. Interestingly, models trained solely on Pyradiomics features demonstrated greater robustness, with the Random Forest model achieving an AUC of 0.78 on the Philips test set.

Conclusions: While radiomics-based ML models show promise for PCa detection, their generalizability across MRI scanners is far from guaranteed. Performance disparities across vendors highlight the critical need

for standardized radiomic feature extraction pipelines to ensure model reliability in real-world clinical applications. Our findings suggest that some radiomics-based models can survive across MRI scanners, but only under carefully controlled conditions—reinforcing the importance of cross-vendor validation in AI-driven diagnostic tools.

Keywords: Machine Learning · Inter-Vendor Reproducibility · Radiomics · Prostate Cancer · Diagnostic tools · Model Reproducibility

1 Introduction

commitment to equitable and patient-centered diagnostic care. Brembilla et al. have underscored model reproducibility as a critical issue in the implementation of AI for prostate MRI [1]. Furthermore, Renard et al., in their analysis of diagnostic imaging variability across different geographic and socioeconomic contexts, have stressed the importance of inclusive research and validation to ensure that ML models are applicable and effective in diverse clinical settings [2].

The bi-parametric prostate MRI protocol known as IMPROD bpMRI has shown promise in reducing unnecessary biopsy procedures while improving the detection of clinically significant PCa [3][4]. This protocol offers a practical MR sequence with reasonable acquisition times and has been designed for ease of implementation across different clinical settings [5][6][3][4]. Its consistent use across various MRI devices and vendor platforms allows for the evaluation of reproducibility in contexts where the imaging protocol remains constant, potentially yielding more reliable and accurate outcomes than protocols with variable acquisition parameters.

The present study addresses these concerns by examining the reproducibility of ML models in assessing tumor aggressiveness from MRI scans across MRI systems from multiple vendors. Through systematic feature selection and the application of an extensive suite of evaluation metrics, this work seeks to demonstrate the feasibility of reproducibility in ML-driven prostate cancer diagnostics and to promote their integration into routine clinical workflows. To this end, two conventional ML models were evaluated using two open-source radiomic feature extraction toolkits—MRCradiomics (https://github.com/haanme/MRCRadiomics) and Pyradiomics (https://github.com/AIM-Harvard/pyradiomics)—to assess their performance across datasets acquired using identical and distinct MRI devices. The study utilized T2-weighted MRI images, which are widely used in PCa imaging and offer clinically acceptable acquisition durations. This approach is intended to support the broader clinical acceptance of AI-enhanced diagnostics by ensuring their precision, consistency, and reliability—qualities that are essential for clinical trust and improved patient outcomes.

2 Materials and Methods

2.1 MRI Data Collection

We used imaging data of men with a clinical suspicion of prostate cancer (PCa) who have been enrolled in prospective, registered, and completed clinical trials: IMPROD (NCT01864135; http://mrc.utu.fi/mri/improd), MULTI-IMPROD (NCT01864135; http://mrc.utu.fi/mri/multi-improd), PROMANEG (NCT02388126), and FLUCIPRO (NCT02002455), along with prostate cancer datasets previously utilized in [5] and [6]. The trials were approved by the Institutional Review Board (IRB), and each enrolled participant provided written informed consent. In total, a pooled cohort of 637 men with a clinical suspicion of PCa were included in the study. The entire imaging protocol included both Diffusion Weighted Imaging using the IMPROD protocol [3] [4], with shimming and calibration, and had an average duration of 15–17 minutes per patient. While the IMPROD bpMRI protocol consisted of optimized T2-weighted (axial and sagittal) and three separate Diffusion Weighted Imaging (DWI) acquisitions, this study used only axial T2-weighted images to assess model performance in scenarios where partial imaging is obtained to reduce MRI acquisition time. All datasets were scanned using the same MRI device (Siemens MAGNETOM Verio 3T), while a subset of one dataset was acquired using a different vendor's MRI device (Philips Ingenia 3T).

2.2 Image Data Post-processing

Prostate cancer aggressiveness was graded using Gleason Grade Groups (GGG) [7], based on samples obtained either through systematic or targeted biopsy. Radiomic feature extraction was then performed using the Pyradiomics package [8] and the MRCradiomics package, developed for repeatable radiomics analyses [9]. We extracted a total of 12,693 radiomic features (11,075 from Pyradiomics and 1,618 from MRCradiomics) from the annotated lesions of each subject. The dataset was split into training, validation, and independent test sets with proportions of 54.6%, 13.6%, and 31.8%, respectively.

2.3 Feature Selection

Feature selection using the Maximum Relevance Minimum Redundancy (MRMR) method was performed to avoid redundancy and to select features that contribute most significantly to the predictive performance of the model, while minimizing overlap among the selected variables [10]. Initially, the top 40 features were automatically selected by the MRMR algorithm and subsequently subjected to univariate analysis to evaluate their individual discriminative power on the training and validation datasets. This process led to the refinement of the feature set to 14 variables demonstrating the highest predictive value. These features were not only individually significant but also collectively contributed to enhanced model precision, thereby supporting the goal of ensuring reproducibility across different MRI systems. The final selection of these features was made under a

commitment to a purely data-driven approach, ensuring that only those features which robustly and reliably reflect biologically relevant characteristics of prostate cancer were retained [11][10]. The MRMR algorithm implemented from the scikit-learn-contrib package (version 0.2.8) was employed to identify the top 14 features used in subsequent analyses, with the number of features selected based on their performance on both the training and validation datasets. Feature selection was applied separately to radiomic features extracted using MRCradiomics and Pyradiomics, as well as jointly, resulting in three distinct sets of selected radiomic features used for comparative evaluation.

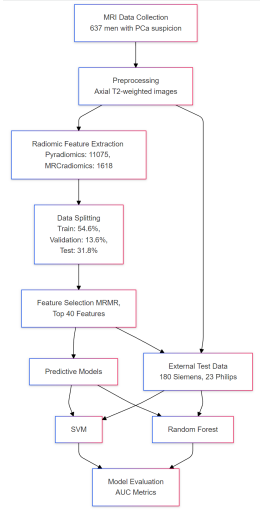


Fig. 1: Flowchart of the study.

2.4 Predictive ML Models

In our experimentation, we divided the cohort of 637 participants into three distinct groups: 434 cases were allocated for training and internal validation, using an 80–20 split to support effective model learning and comprehensive self-assessment. The remaining cases were reserved for external evaluation: 180 subjects were scanned using the same MRI vendor (Siemens), and 23 subjects were

scanned using a different vendor's MRI device (Philips). This setup allowed for a rigorous assessment of model generalizability across heterogeneous imaging platforms.

We evaluated two widely adopted machine learning techniques by constructing both a Support Vector Machine (SVM) model and a Random Forest model. For the former, given SVM's strength in binary classification tasks [12], we handled missing values through median imputation, a step that is particularly important to maintain classifier stability and performance when dealing with high-dimensional radiomic features [13]. For the latter, a Random Forest model was implemented. Both models were trained using the same data splits and the identical feature selection pipeline, ensuring a consistent and controlled comparative evaluation.

3 Result

The selected features for Pyradiomics, MRCradiomics, and their combination (referred to as Pyradiomics+MRCradiomics) are detailed in Tables 1, 2, and 3 (Appendix), respectively.

3.1 Pyradiomics Features: Consistency and Robustness

Models trained exclusively on Pyradiomics-derived features demonstrated greater stability and superior predictive performance across different scanner vendors. The Random Forest model achieved an AUC of 0.68 (90% CI: [0.61, 0.74]) on the Multi-IMPROD dataset and an AUC of 0.78 (90% CI: [0.61, 0.94]) on the Philips dataset. Similarly, the SVM model trained solely on Pyradiomics features yielded an AUC of 0.77 (90% CI: [0.59, 0.92]) on the Philips test set. However, its performance was diminished on the Multi-IMPROD dataset, with an AUC of 0.60 (90% CI: [0.53, 0.66]), indicating reduced consistency within the same scanner vendor context compared to Random Forest.

3.2 MRCradiomics Features: Variability in Generalization

The predictive performance of models trained using MRCradiomics-derived features exhibited greater variability across datasets. The Random Forest model achieved an AUC of 0.68 (90% CI: [0.63, 0.74]) on the Multi-IMPROD dataset but showed limited robustness on the Philips test set, where the AUC declined to 0.56 (90% CI: [0.40, 0.73]). The SVM model produced an AUC of 0.73 (90% CI: [0.67, 0.78]) on the Multi-IMPROD dataset, but only 0.53 (90% CI: [0.36, 0.71]) on the Philips test set. These findings suggest that models trained on MRCradiomics features may be more sensitive to vendor-induced variability.

3.3 Model Performance on Combined Features

When leveraging the combined feature set from Pyradiomics and MRCradiomics, the SVM model achieved a promising AUC of 0.74 (90% CI: [0.68, 0.79]) on the

Multi-IMPROD dataset, highlighting its effectiveness in utilizing heterogeneous radiomic feature sets within a uniform scanning environment. However, its performance deteriorated on the Philips test set, where it achieved a markedly lower AUC of 0.35 (90% CI: [0.22, 0.49]), emphasizing its vulnerability to scanner-specific variation. Similarly, the Random Forest model achieved an AUC of 0.73 (90% CI: [0.68, 0.79]) on the Multi-IMPROD dataset, but its generalizability was challenged on the Philips test set, where performance declined to an AUC of 0.60 (90% CI: [0.44, 0.76]).

The results highlight substantial variability in model performance attributable to scanner-specific differences. Pyradiomics-derived features proved to be more reliable and generalizable across scanner vendors, particularly with the Random Forest model, which attained the highest AUC on the Philips dataset. In contrast, models trained using MRCradiomics-derived features demonstrated increased susceptibility to scanner-induced discrepancies, thereby limiting their cross-platform generalizability.

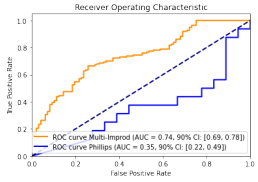


Fig. 2: Result of SVM trained over Pyradiomics and MRCradiomics

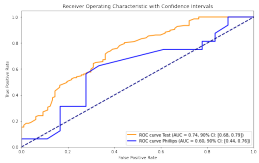


Fig. 3: Results of Random Forest trained over Pyradiomics and MRCradiomics

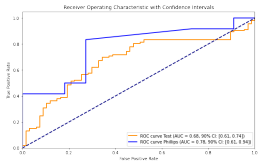


Fig. 4: Result of Random Forest trained over Pyradiomics

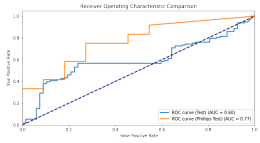


Fig. 5: Results of SVM trained over Pyradiomics

4 Discussion

In this comprehensive study, we conducted an in-depth evaluation of machine learning (ML) models for predicting prostate cancer aggressiveness, leveraging advanced radiomic features extracted using the Pyradiomics and MRCradiomics packages. Our analysis demonstrates the potential of combining features from both packages to enhance predictive performance, while also revealing critical challenges associated with cross-vendor variability and generalizability. Methodological rigor was ensured through extensive hyperparameter optimization using RandomizedSearchCV for Random Forest and GridSearchCV for SVM, with the area under the ROC curve (AUC) employed as the primary evaluation criterion [14] [15] [16]. This framework reinforces our objective of developing ML-based clinical decision support systems (CDS) with enhanced precision and robustness [14] [17].

The SVM model, a robust classifier designed to optimize decision boundaries in high-dimensional spaces [18] [19], exhibited strong performance on the Multi-IMPROD dataset, achieving an AUC of 0.74. This result underscores its ability to effectively utilize heterogeneous radiomic features [20]. However, its sensitivity to imaging platform variability became apparent as performance declined to an AUC of 0.60 on the Philips test set [21]. Similarly, the Random Forest model, known for its ensemble-based approach that mitigates overfitting and improves generalization [22], achieved an AUC of 0.73 on the Multi-IMPROD dataset. Yet, like the SVM, its performance declined to 0.60 on the Philips test set, highlighting scanner-induced variability as a persistent challenge [23]. Despite this, Random Forest demonstrated superior interpretability compared to SVM, especially when trained on Pyradiomics features, achieving the highest

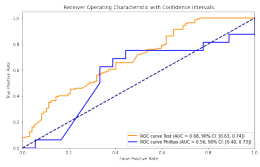


Fig. 6: Result of Random Forest trained over MRCradiomics.

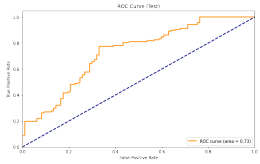


Fig. 7: Results of SVM trained over MRCradiomics and tested on Multi-Improd Dataset.

AUC of 0.78 on the Philips test set [24]. This advantage can be attributed to the model's intrinsic capacity to rank feature importance and extract interpretable insights from complex datasets [25]. In contrast, SVM's dependence on kernel-based transformations limits model transparency [26], an essential consideration in clinical contexts where interpretability is as important as predictive performance [18].

Our study highlights the significance of radiomic feature selection and representation in determining model generalizability. The combined use of Pyradiomics and MRCradiomics features produced competitive AUCs on the Multi-IMPROD dataset but revealed substantial performance variability across scanners. This emphasizes the necessity of validating radiomic signatures across diverse imaging platforms to improve reproducibility [14] [15]. Notably, Pyradiomics-derived features consistently outperformed MRCradiomics features, particularly on the Philips test set. These findings suggest that Pyradiomics features may encapsulate more reproducible and clinically informative characteristics, making them a more suitable choice for developing generalizable CDS systems [14] [15].

The observed disparities in model performance between the Multi-IMPROD and Philips test sets reveal critical sources of heterogeneity, arising from differences in imaging systems and dataset sizes. The Multi-IMPROD dataset, obtained using a Siemens MRI scanner, contained 180 cases, offering a sufficiently large and diverse sample for model evaluation. In contrast, the Philips test set included only 23 cases, limiting the statistical power and model generalizability [23]. Moreover, differences in imaging resolution, contrast parameters, and

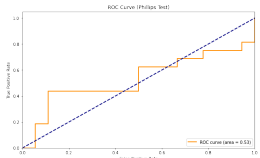


Fig. 8: Results of SVM trained over MRCradiomics and tested on Philips Dataset.

acquisition protocols between Siemens and Philips scanners likely contributed to variations in the extracted radiomic features, subsequently affecting model performance [21]. Addressing these inter-scanner discrepancies is essential for advancing ML models capable of consistent performance across clinical environments [15] [14].

While this study provides valuable insights into the reproducibility of radiomics-based ML models, several limitations remain. First, the limited sample size of the Philips test set restricts the external validity of our findings. Future research should incorporate larger, multi-institutional datasets to enhance statistical robustness and reduce uncertainty due to sample variance. Second, although Pyradiomics features showed higher cross-scanner reproducibility, the underlying causes of this robustness warrant further exploration. Investigating feature harmonization techniques such as ComBat or deep feature alignment methods may yield further improvements. Third, this study primarily explored supervised learning algorithms. Future work could benefit from incorporating unsupervised, semi-supervised, or transfer learning approaches to enable adaptation to new imaging contexts with limited annotations. Additionally, attention-based architectures such as Vision Transformers may provide improved performance and interpretability for modeling scanner-induced variations [15].

5 Conclusion

In this study, we have comprehensively examined the efficacy of machine learning models in predicting the aggressiveness of prostate cancer through the utilization of advanced radiomic features extracted using the Pyradiomics and MRCradiomics packages. By exploring a broad spectrum of hyperparameters for Random Forest and SVM algorithms, we sought to optimize model performance through systematic tuning using RandomizedSearchCV and GridSearchCV, with the area under the ROC curve (AUC) employed as the primary evaluation metric. Our findings underscore the potential of combining features from both Pyradiomics and MRCradiomics to enhance predictive performance in healthcare diagnostics, particularly in the context of prostate cancer risk stratification. The evaluation of our models revealed substantial variability in predictive

performance across different MRI platforms, with the Multi-IMPROD dataset (acquired on a Siemens scanner) yielding higher AUC scores compared to the Philips test set. This discrepancy is attributable to both differences in dataset sizes and variations in imaging protocols and hardware configurations between Siemens and Philips scanners. Specifically, the larger and more diverse Multi-IMPROD dataset enabled more effective model training and validation, while the smaller Philips dataset introduced higher variance and reduced generalizability. These findings emphasize the importance of cross-platform validation of radiomic features to ensure the reliability, reproducibility, and clinical applicability of machine learning-based Clinical Decision Support Systems (CDS). By demonstrating the superior predictive performance and cross-scanner robustness of Pyradiomics-derived features, our study highlights their potential as a more stable and reproducible radiomic feature set for model development in heterogeneous imaging environments.

Appendix

Feature Name	Value
log_sigma_1_GLDM_LargeDependence	0.672
log_sigma_2_GLRLM_ShortRunHighLevelEmph	0.651
log_sigma_2_GLRLM_HighGrayLevelRunEmph	0.649
log_sigma_2_GLDM_HighGrayLevelEmph	0.649
log_sigma_1_GLSZM_SizeZoneNonUniformity	0.647
log_sigma_2_GLSZM_HighGrayLevelZoneEmph	0.646
log_sigma_1_GLRLM_ShortRunHighLevelEmph	0.646
wavelet_HLL_GLDM_LargeDependence	0.645
log_sigma_1_GLRLM_HighGrayLevelRunEmph	0.645
log_sigma_1_GLDM_HighGrayLevelEmph	0.644
log_sigma_1_GLCM_Autocorrelation	0.644
diagnostics_Mask_interpolated_Maximum	0.644
original_firstorder_Maximum	0.644
log_sigma_1_firstorder_Range	0.642
original_firstorder_Range	0.641

Table 1: Selected radiomic features extracted with pyradiomics package. Features were selected with MRMR method using T2-weighted images of 434 prostate cancer subjects, their respective descriptions, package names, and AUC performance in validation set.

Feature Name	AUC
0.05_No_corners_ROI	0.679
0.01_No_corners_ROI	0.679
0.05_No_corners_ROI	0.678
0.01_No_corners_ROI	0.677
0.01_No_corners_ROI	0.674
objprops_N_objs	0.673
objprops_Per_IQR_mm	0.671
0.05_No_corners_ROI	0.670
objprops_Int_IQR	0.668
Frangi_objprops_Int_SD	0.667
Frangi_objprops_Int_IQR	0.665
Scharr_objprops_Area_median_mm2	0.665
Range	0.663
Scharr_objprops_Area_mean_mm2	0.663
Hessian_0.025_15.0_objprops_Int_SD	0.658

Table 2: Selected radiomic features extracted with MRCradiomics package. Features were selected with MRMR method using T2-weighted images of 434 prostate cancer subjects, their respective descriptions, package names, and AUC performance in validation set.

Description	Package	AUC
Wavelet LLL gldm DependenceEntropy	PYR	0.697
Wavelet LLL glszm ZoneEntropy	PYR	0.696
Wavelet LLL glcm JointEntropy	PYR	0.694
Original shape LeastAxisLength	PYR	0.690
Log sigma 2.0 mm 3D gldm DependenceEntropy	PYR	0.686
1 mm original gldm DependenceEntropy	PYR	0.682
1 mm Wavelet LHL glcm Idmn	PYR	0.681
1 mm log sigma 3.0 mm 3D glcm DifferenceEntropy	PYR	0.678
1 mm Wavelet LHH gldm DependenceVariance	PYR	0.677
1 mm Wavelet HHL glcm Imc1	PYR	0.677
1 mm Wavelet LHL glcm Idn	PYR	0.676
Harris-Stephens corner-edge b4 ks7 k0.50 Corner density primary	MRCR	0.666
Harris-Stephens corner-edge b4 ks7 k0.50 Corner density mean	MRCR	0.662
Scharr filtered object properties: Number of objects	MRCR	0.661

Table 3: Selected radiomic features extracted with pyradiomics(PYR) and MRCradiomics(MRCR) packages. Features were selected with MRMR method using T2-weighted images of 434 prostate cancer subjects, their respective descriptions, package names, and AUC performance in validation set. (PY: Pyradiomics and MRC: MRCradiomics)

References

1. Giorgio Brembilla, Francesco Giganti, Harbir Sidhu, Massimo Imbriaco, Sue Mallett, Armando Stabile, Alex Freeman, Hashim U Ahmed, Caroline Moore, Mark

Emberton, et al. Diagnostic accuracy of abbreviated bi-parametric mri (a-bpmri) for prostate cancer detection and screening: a multi-reader study. *Diagnostics*, 12(2):231, 2022.
2. Félix Renard, Soulaimane Guedria, Noel De Palma, and Nicolas Vuillerme. Variability and reproducibility in deep learning for medical image segmentation. *Scientific Reports*, 10(1):13724, 2020.
3. O. Ettala, I. Jambor, I.M. Perez, M. Seppänen, A. Kaipia, H. Seikkula, K.T. Syvänen, P. Taimen, J. Verho, A. Steiner, and J. Saunavaara. Individualised non-contrast mri-based risk estimation and shared decision-making in men with a suspicion of prostate cancer: protocol for multicentre randomised controlled trial (multi-improd v. 2.0). *BMJ open*, 12(4):e053118, 2022.
4. I. Jambor, P.J. Boström, P. Taimen, et al. Novel biparametric mri and targeted biopsy improves risk stratification in men with a clinical suspicion of prostate cancer (improd trial). *J Magn Reson Imaging*, 46(4):1089–1095, 2017.
5. I. Jambor, E. Kähkönen, P. Taimen, H. Merisaari, J. Saunavaara, K. Alanen, B. Obsitnik, H. Minn, V. Lehotska, and H.J. Aronen. Prebiopsy multiparametric 3t prostate mri in patients with elevated psa, normal digital rectal examination, and no previous biopsy. *Journal of Magnetic Resonance Imaging*, 41(5):1394–1404, 2015.
6. E. Kähkönen, I. Jambor, J. Kemppainen, K. Lehtiö, T.J. Grönroos, A. Kuisma, P. Luoto, H.J. Sipilä, T. Tolvanen, K. Alanen, and J. Silén. In vivo imaging of prostate cancer using [68ga]-labeled bombesin analog bay86-7548. *Clinical cancer research*, 19(19):5434–5443, 2013.
7. S. Loeb, Y. Folkvaljon, D. Robinson, I.F. Lissbrant, L. Egevad, and P. Stattin. Evaluation of the 2015 gleason grade groups in a nationwide population-based cohort. *European urology*, 69(6):1135–1141, 2016.
8. J.J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G. Beets-Tan, J.C. Fillion-Robin, S. Pieper, and H.J. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.
9. H. Merisaari, P. Taimen, R. Shiradkar, O. Ettala, M. Pesola, J. Saunavaara, P.J. Boström, A. Madabhushi, H.J. Aronen, and I. Jambor. Repeatability of radiomics and machine learning for dwi: Short-term repeatability study of 112 patients with prostate cancer. *Magnetic resonance in medicine*, 83(6):2293–2309, 2020.
10. Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.
11. Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.
12. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.
13. Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.
14. Xiaoyang Qi, Kai Wang, B. Feng, Xing-Ai Sun, Jie Yang, Zhengbiao Hu, Maoliang Zhang, Cheng Lv, Liyuan Jin, Lingyan Zhou, Zheng-ping Wang, and Jincao Yao. Comparison of machine learning models based on multi-parametric magnetic resonance imaging and ultrasound videos for the prediction of prostate cancer. *Frontiers in Oncology*, 13, 2023.

15. E. Gresser, B. Schachtner, Anna Theresa Stüber, O. Solyanik, Andrea Schreier, T. Huber, M. Froelich, G. Magistro, A. Kretschmer, C. Stief, J. Ricke, M. Ingrisch, and Dominik Nörenberg. Performance variability of radiomics machine learning models for the detection of clinically significant prostate cancer in heterogeneous mri datasets. *Quantitative Imaging in Medicine and Surgery*, 12:4990–5003, 2022.
16. Hailang Liu, K. Tang, E. Peng, Liang Wang, D. Xia, and Zhiqiang Chen. Predicting prostate cancer upgrading of biopsy gleason grade group at radical prostatectomy using machine learning-assisted decision-support models. *Cancer Management and Research*, 12:13099–13110, 2020.
17. Kai Wang, Pei-Ling Chen, Bojian Feng, Jing Tu, Zhengbiao Hu, Maoliang Zhang, Jie Yang, Y. Zhan, Jincao Yao, and Dong-Guo Xu. Machine learning prediction of prostate cancer from transrectal ultrasound video clips. *Frontiers in Oncology*, 12, 2022.
18. Chengquan Huang, L Davis, and J Townshend. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23:725–749, 2002.
19. Shijin Wang, A Mathew, Yan Chen, L Xi, Lin Ma, and Jay Lee. Empirical analysis of support vector machine ensemble classifiers. *Expert Syst. Appl.*, 36:6466–6476, 2009.
20. E Niaf, R Flamary, O Rouvière, C Lartizien, and S Canu. Kernel-based learning from both qualitative and quantitative labels: Application to prostate cancer diagnosis based on multiparametric mr imaging. *IEEE Transactions on Image Processing*, 23:979–991, 2014.
21. Antonio Candelieri and D Conforti. A hyper-solution framework for svm classification: Application for predicting destabilizations in chronic heart failure patients. *The Open Medical Informatics Journal*, 4:136–140, 2010.
22. Archana Gunakala and Afzal Hussain Shahid. A comparative study on performance of basic and ensemble classifiers with various datasets. *Applied Computer Science*, 2023.
23. G Madzarov, D Gjorgjevikj, and I Chorbev. A multi-class svm classifier utilizing binary decision tree. *Informatica (Slovenia)*, 33:225–233, 2009.
24. Weili Jiang, Zhenhua Chen, Y Xiang, Dangguo Shao, Lei Ma, and Junpeng Zhang. Ssem: A novel self-adaptive stacking ensemble model for classification. *IEEE Access*, 7:120337–120349, 2019.
25. Prudhvi K Gurram and H Kwon. Sparse kernel-based ensemble learning with fully optimized kernel parameters for hyperspectral classification problems. *IEEE Transactions on Geoscience and Remote Sensing*, 51:787–802, 2013.
26. B Haasdonk. Feature space interpretation of svms with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:482–492, 2005.

Jatin Chaudhary and Dipak Nidhi and Jukka Heikkonen and Haari Merisaari and Rajiv Kanth
Super Level Sets and Exponential Decay: A Synergistic Approach to Stable Neural Network Training

V

Journal of Artificial Intelligence Research, Accepted for Publication

Super-level Sets and Exponential Decay: A Synergistic Approach to Stable Neural Network Training

JATIN CHAUDHARY*, University of Turku, Finland
DIPAK NIDHI, University of Turku, Finland
JUKKA HEIKKONEN, University of Turku, Finland
HARRI MERISAARI, University of Turku, Finland
RAJEEV KANTH, Savonia University of Applied Sciences, Finland

This paper presents a theoretically grounded optimization framework for neural network training that integrates exponentially decaying learning rates with Lyapunov-based stability analysis. We develop a dynamic learning rate algorithm and prove that it induces connected and stable descent paths through the loss landscape by maintaining the connectivity of super-level sets $S_\lambda = \{\theta \in \mathbb{R}^n : \mathcal{L}(\theta) \geq \lambda\}$. Under the condition that the Lyapunov function $V(\theta) = \mathcal{L}(\theta)$ satisfies $\nabla V(\theta) \cdot \nabla \mathcal{L}(\theta) \geq 0$, we establish that these super-level sets are not only connected but also equiconnected across epochs, providing uniform topological stability. We further derive convergence guarantees using a second-order Taylor expansion and demonstrate that our exponentially scheduled learning rate with gradient-based modulation leads to a monotonic decrease in loss. The proposed algorithm incorporates this schedule into a stability-aware update mechanism that adapts step sizes based on both curvature and energy-level geometry. This formulation contributes to the theoretical foundations of neural optimization by formalizing the role of topological structure in convergence dynamics and offers a provably stable and adaptive training scheme for high-dimensional and non-convex learning problems.

JAIR Associate Editor: Prof. Bo Han

JAIR Reference Format:
Jatin Chaudhary, Dipak Nidhi, Jukka Heikkonen, Harri Merisaari, and Rajeev Kanth. 2025. Super-level Sets and Exponential Decay: A Synergistic Approach to Stable Neural Network Training. *JAIR* 1, Article 6 (August 2025), 19 pages. DOI: 10.1613/jair.1.xxxxx

1 Introduction

There has been significant progress towards the development and deployment of neural network models. The deployment of a neural network model demands, high accuracy and precision, and hyperparameter optimization plays an important role towards building such a model. The researchers' community has been actively analyzing learning rates, and loss functions, to make the network more stable across datasets, and prevent overfiting [28][5][18]. Optimizing neural networks involves minimizing a complex and often non-convex loss function over a high-dimensional parameter space. These non-convex landscapes present significant challenges as gradient-based methods can become trapped in suboptIn the following sections, we delve into the mathematical foundations

*Corresponding Author.

Authors' Contact Information: Jatin Chaudhary, jatin.chaudhary@utu.fi, ORCID: 0000-0002-4139-5315, University of Turku, Turku, Finland; Dipak Nidhi, ORCID: 0000-0002-1040-5007, dipak.nidhi@utu.fi, University of Turku, Turku, Finland; Jukka Heikkonen, ORCID: 0000-0002-2468-5708, jukka.heikkonen@utu.fi, University of Turku, Turku, Finland; Harri Merisaari, ORCID: 0000-0002-8515-5399, haanme@utu.fi, University of Turku, Turku, Finland; Rajeev Kanth, ORCID: 0000-0003-1109-1211, Rajeev.Kanth@savonia.fi, Savonia University of Applied Sciences, Kuopio, Finland.

DOI: 10.1613/jair.1.xxxxx

that link dynamic learning rates with super-level sets, crucial loss function for understanding stability and convergence in neural network training. We will explore how adaptive learning rates, particularly those with exponential decay, systematically influence the optimization landscape. This discussion aims to bridge theoretical insights with practical strategies, enhancing both the efficacy and understanding of neural network training in local minima or saddle points [6]. Despite these difficulties, substantial progress has been made in understanding and enhancing optimization trajectories in neural networks. Recent theoretical advancements have highlighted concepts like 'loss landscape smoothing' and adaptive gradient methods,' indicating that certain learning rate configurations can improve optimization conditions [16].

Neural network training presents multiple challenges, particularly in optimizing the learning rate, managing the loss function, ensuring stability, and preventing overfitting. The learning rate is a critical parameter that dictates the step size during gradient descent. An inappropriate learning rate can lead to slow convergence or even divergence. The loss function, which measures the discrepancy between predicted and actual outputs, often has a complex landscape that can trap optimization algorithms in local minima [22]. Stability is another crucial aspect, as unstable training can lead to erratic updates and poor model performance. Overfitting, where the model performs well on training data but poorly on unseen data, remains a persistent problem. Existing solutions include adaptive learning rates and regularization techniques, but they often fall short in ensuring consistent stability and avoiding overfitting [23]. Our study addresses these issues by proposing a novel approach that integrates dynamic learning rates with stability principles from control theory.

Our primary contribution is the development of an algorithm that dynamically adjusts the learning rate using an exponential decay model, integrated with principles from Lyapunov stability [4]. This approach ensures consistent convergence by maintaining the connectivity of super-level sets of the loss function. We demonstrate that these super-level sets remain connected under our algorithm, preventing the optimization process from becoming trapped in poor local minima and ensuring stable descent paths [8]. This connectedness facilitates smoother transitions across the loss landscape, enhancing training dynamics and generalization capabilities. By embedding these concepts into our algorithmic framework, we achieve more stable and efficient optimization, addressing common challenges such as overfitting and instability. This work not only advances theoretical understanding but also provides a foundation for practical applications in neural network training, paving the way for further research into dynamic learning rate adjustments and their impact on training stability and efficacy [27].

In the following sections, we present the mathematical foundations, and further link dynamic learning rates with super-level set loss function, crucial for understanding stability and convergence in neural network training. We will explore how adaptive learning rates, particularly those with exponential decay, systematically influence the optimization landscape. We discuss the stability of using adaptive learning rates with super-level set loss function so to solidify our claims.

2 Mathematical Underpinnings

The super-level sets $S_\lambda = \{\mathbf{x} \in \mathbb{R}^n : L(\mathbf{x}) \geq \lambda\}$ reveal important stability and convergence properties for gradient-based optimization methods [15]. An exponentially decaying learning rate, defined by $\eta(t) = \eta_0 e^{-\alpha t}$, where η_0 is the initial rate and α a positive decay constant, is beneficial. It allows for quick initial progress by using a higher initial rate, guiding the optimizer towards important areas quickly [12]. As training proceeds, this rate gradually decreases, allowing for more precise adjustments and preventing common issues like overshooting minima [10]. This dynamic rate adjustment, when coupled with the structure of super-level sets, offers insights into the training's stability by ensuring the optimization path remains connected and stable through the topology of the landscape [23]. By adopting a Lyapunov function $V(\mathbf{x})$ that decreases along these paths, we enforce stability and keep the system's energy diminishing, keeping the optimization within stable parameter regions [4]. Together,

Table 1. Summary of Notations Used Throughout the Paper

Symbol	Description
θ	Trainable parameters of the neural network
$\mathcal{L}(\theta)$	Global Loss function (cross-entropy unless stated)
f	Per-sample loss
$V(\theta)$	Lyapunov function (set to $\mathcal{L}(\theta)$ for analysis)
$\nabla\mathcal{L}(\theta)$	Gradient of loss with respect to parameters
$\alpha(t)$	Exponentially decaying learning rate at epoch t
η_t	Norm-scaled dynamic learning rate
S_λ	Superlevel set: $\{\theta : \mathcal{L}(\theta) \geq \lambda\}$
g_t	Gradient at iteration t (i.e., $g_t = \nabla_\theta \mathcal{L}(\theta_t)$)
x_t	Iteration-dependent point in parameter space (used interchangeably with θ_t)
Δ_t	Accumulated shift from reference point x_{ref}
m_t, v_t, r_t	First and second moment estimates for gradient, variance, and rate
β	Vector Moment decay
λ	Hyperparameter controlling exponential decay
x_{BASE}	Base model initialization

these elements create a robust framework that deepens our understanding of the dynamics in neural network training and highlights the significance of careful tuning of hyperparameters in managing complex optimization scenarios [27].

To understand the concept better, consider a ball that rolls down a hilly terrain towards a valley, representing the minimum of a loss landscape. Initially, the ball is given a strong push (high initial learning rate η_0) allowing it to quickly descend from higher elevations (higher loss values in super-level sets S_λ). Each super-level set corresponds to a range of elevations where the ball's potential energy (analogous to the loss value in the neural network) remains above a certain threshold λ. As the ball descends from higher altitudes to lower ones, it transitions from one super-level set to another, each with decreasing minimum energy thresholds. As it approaches the valley, the slope (gradient) lessens and so does the ball's speed due to the exponential decay of the push force ($\eta(t) = \eta_0 e^{-\alpha t}$), preventing it from overshooting the valley. This gradual slowing is critical as it ensures that the ball can finely adjust its path to settle in the deepest part of the valley, analogous to achieving the most optimal parameters in a neural network training scenario. This model demonstrates how the dynamic learning rate and the structure of the super-level sets interact, ensuring that the optimization path remains stable and connected throughout the descent, analogous to how the ball consistently follows a path that leads it towards the valley without getting stuck or veering off course.

3 Fundamental Concepts

The parameter vector θ, has the network's weights and biases, and is an integral part for the network's learning, as it is meticulously adjusted to minimize divergences between predicted outputs and actual targets [21]. This adjustment process is governed by the learning rate $\alpha(t)$, a parameter that determines the step size within the parameter space during optimization, thus directly influencing convergence quality [17]. The gradient of the loss function, $\nabla_\theta \mathcal{L}(\theta)$, serves as the navigational guide for updating parameters, towards optimal solutions [29]. The interplay between the learning rate and the gradient is vital for maintaining systematic progression and ensuring that the training remains on a stable and effective path [14]. This setup forms the backbone of our approach to

enhancing neural network training, laying the groundwork for a deeper exploration of optimization dynamics mathematically [3].

3.1 Mathematical Draw Outs

Behind our study is a probabilistic model that views the neural network as an intricate function approximating the conditional probability distribution $P(Y \mid X; \theta)$. In classification tasks, this relationship is mathematically expressed through the softmax function:

$$P(Y = c \mid X; \theta) = \frac{\exp\left(f_c(X; \theta)\right)}{\sum_{j=1}^{C} \exp\left(f_j(X; \theta)\right)}, \tag{1}$$

where $f_c(X; \theta)$ represents the network output for class c, and C denotes the total number of classes [1]. This formulation is essential in demonstrating how our model probabilistically classifies input data into defined output classes.

Building on this framework, we derive a likelihood function reflecting the probability of observing our training dataset $\mathcal{D} = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$ under the model parameters θ :

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{m} P\left(y^{(i)} \mid x^{(i)}; \theta\right). \tag{2}$$

This likelihood function for quantifying how well the model aligns with empirical data, setting the stage for parameter optimization via Bayesian inference [13].

Incorporating Bayesian principles, we consider the posterior probability of the parameters θ given the data $\mathcal{D}$, calculated as follows:

$$P(\theta \mid \mathcal{D}) \propto \mathcal{L}(\theta; \mathcal{D}) P(\theta), \tag{3}$$

where $P(\theta)$ denotes the prior distribution over the parameters [2]. This Bayesian framework facilitates a comprehensive parameter optimization strategy, harmonizing empirical data adaptation with existing parameter knowledge.

The culmination of this probabilistic modeling leads to the optimization phase within a gradient descent framework, where our methodology involves iteratively minimizing the negative log-posterior:

$$-\log P(\theta \mid \mathcal{D}) = -\log \mathcal{L}(\theta; \mathcal{D}) - \log P(\theta) + \text{ const.} \tag{4}$$

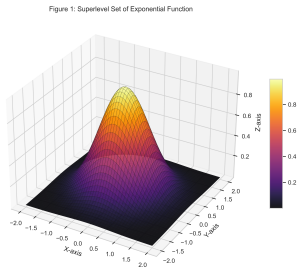
Here, the gradient descent update rule is critical:

$$\theta_{t+1} = \theta_t - \alpha(t) \nabla_\theta \left[-\log P\left(\theta_t \mid \mathcal{D}\right)\right], \tag{5}$$

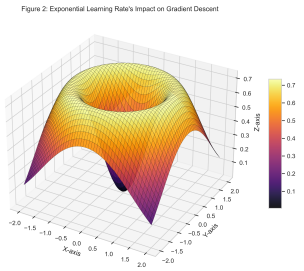
where $\alpha(t)$ is the learning rate, dynamically adapting to ensure efficient convergence and stability of the model [17].

Importantly, the dynamic adjustment of $\alpha(t)$ profoundly impacts the topology of the loss function's super-level sets $S_\lambda = \{\theta \in \mathbb{R}^n : \mathcal{L}(\theta) \geq \lambda\}$, which are instrumental in understanding the stability and connectivity of the optimization landscape [6]. By ensuring that these sets remain connected, the algorithm promotes a smoother and more stable descent toward the global minima, effectively navigating the complex, high-dimensional parameter spaces typical of deep learning tasks[33].

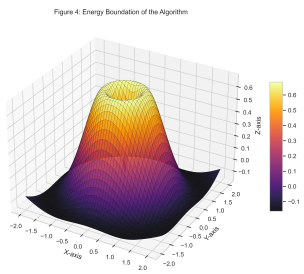
This integration of probabilistic modeling, Bayesian inference, and gradient optimization leverages the theoretical insights into super-level sets to enhance the practical outcomes of neural network training. This approach ensures both theoretical robustness and empirical efficacy, highlighting our model's capacity to navigate and optimize within intricate, probabilistically defined landscapes.



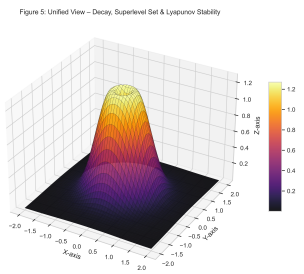
Superlevel Set of Exponential Function



Exponential Learning Rate's Impact



Energy Boundation of the Algorithm



Unified View: Decay, Superlevel Set & Stability

Fig. 1. **Geometric and Energetic Intuition Behind the Proposed Optimization Framework.** This 2×2 composite figure presents key geometric and dynamic principles central to our theoretical framework. **(a)** The surface of $f(x, y) = \exp(-x^2-y^2)$ depicts the loss landscape, where level contours form smoothly connected superlevel sets S_λ. **(b)** The exponential decay in learning rate $\alpha(t)$ modulates the influence of steep curvature regions, effectively scaling updates in line with gradient norm. **(c)** The Lyapunov-inspired energy profile demonstrates how bounded update energy suppresses divergence, stabilizing convergence across iterations. **(d)** Finally, the unified visualization integrates exponential decay, superlevel containment, and energy boundation showing how our algorithm balances descent efficiency with theoretical stability. Together, these views provide a visual bridge between our mathematical claims (Sections 4–6) and their algorithmic implications.

3.2 Exponentially Decaying Learning Rate

The formulation of the Exponentially Decaying Learning Rate (derivation in the supplementary) given by

$$\frac{d\alpha}{dt} = -\alpha_0 \beta e^{-\beta t}, \tag{6}$$

influences the topology of the loss function's super-level sets $S_\lambda = \{\theta \in \mathbb{R}^n : \mathcal{L}(\theta) \geq \lambda\}$. The dynamically adjusted learning rate ensures that these sets remain connected, supporting a stable and cohesive optimization

trajectory [12]. Within the gradient descent framework, this leads to an adapted parameter update rule

$$\theta_{t+1} = \theta_t - \alpha_0 e^{-\beta t} \nabla_\theta \left[-\log P\left(\theta_t \mid \mathcal{D}\right) \right], \tag{7}$$

effectively illustrating the integration of an exponential decay learning rate within the gradient descent mechanism [17]. This methodical approach not only enhances the theoretical underpinnings of our optimization strategy but also significantly boosts its practical efficacy. By marrying the theoretical concepts of exponential decay with gradient descent, our approach fosters training dynamics that effectively navigate the complex, high-dimensional spaces typical of deep learning tasks [23]. This novel integration offers a rigorous, theoretically informed enhancement to the conventional training paradigms, ensuring that both the stability and the efficiency of the learning process are maximized [8].

3.2.1 Dynamic Cost Function. In our study, we refined our dynamic cost function to adeptly integrate principles from statistical learning theory, with an emphasis on addressing class imbalances and evolving training requirements. The empirical risk, $R_{\text{emp}}(\theta)$, is meticulously calculated as

$$\frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i; \theta)), \tag{8}$$

where we incorporate class weights w_c to balance the influence of underrepresented classes, resulting in

$$R_{\text{emp}}^{\text{cw}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} w_{y_i} L(y_i, f(x_i; \theta)) \tag{9}$$

This weighting corrects training biases, enhancing model fairness and accuracy particularly in scenarios with skewed class distributions [24]. To manage outliers and enhance robustness, we introduce a robustness parameter ρ, which modifies the loss contribution based on the confidence in data point correctness:

$$R_{\text{robust}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \rho(y_i, x_i) w_{y_i} L(y_i, f(x_i; \theta)) \tag{10}$$

[35]. Regularization is integral to this framework, implemented through $\Omega(\theta)$, employing either L_1 or L_2 regularization to mitigate overfitting. The regularized empirical risk is articulated as

$$R_{\text{emp}}^{\text{reg}}(\theta) = R_{\text{robust}}(\theta) + \lambda \Omega(\theta) \tag{11}$$

[11]. Our dynamic cost function is characterized by a temporal modulation factor $\gamma(t) = 1 + \kappa e^{-\delta t}$, which strategically transitions from aggressive initial learning to increased regularization as training advances [30]. This modulation ensures the learning rate evolves with the model's needs, reducing to prevent overfitting as the model refines its parameters. The gradient of the loss function, $\nabla_\theta \mathcal{L}(\theta)$, directs parameter updates and is essential for navigating both the explicit regions, where gradients are large and clear, facilitating straightforward descent steps, and the implicit regions, where gradients may vanish, requiring the adaptive $\gamma(t)$ and robustness enhancements to maintain meaningful and stable updates [17]. For instance, in scenarios with imbalanced datasets, class weights w_c counteract the bias toward predominant classes, and $\gamma(t)$'s increasing regularization later in training smooths the model's fit to emphasize generalization. This framework,

$$\mathcal{J}_{\text{dynamic}}\left(\theta; \mathcal{D}, t\right) = \gamma(t) \mathcal{J}_{\text{reg}}\left(\theta; \mathcal{D}\right), \tag{12}$$

not only deepens our understanding of dynamic learning rate mechanisms but also fosters a coherent and stable optimization process, adaptable to complex data landscapes and advancing adaptive machine learning methodologies [25].

3.3 Gradient Descent

Integrating level set dynamics into the gradient descent framework is proposed to navigate the complex topology of the loss function more efficiently. While traditional gradient descent updates parameters iteratively with the rule $\theta_{t+1} = \theta_t - \alpha(t)\nabla_\theta \mathcal{L}(\theta_t)$, where $\alpha(t) = \alpha_0 e^{-\beta t}$ is an exponentially decaying learning rate, emerging research suggests enhancements to this approach to address its limitations in stability and adaptability. Zhang et al. (2019) propose an Adaptive Exponential Decay Rate (AEDR), which dynamically adjusts the decay rate based on moving averages of gradients, thus offering a more responsive adaptation to the learning needs over different training phases and potentially leading to improved convergence rates [34].

Further, Mishra and Ghosh (2019) highlight the advantages of a variable gain gradient descent, which modulates the learning rate based on error metrics and system states to enhance both the convergence speed and stability, suggesting a potential direction for refining level set dynamics integration [26]. Additionally, the link between generalization and dynamical robustness presented by Kozachkov et al. (2023) through Riemannian contraction indicates that ensuring algorithmic stability through the optimization dynamics could directly influence generalization performance, advocating for a deeper theoretical integration of level set dynamics with gradient descent methods [20].

To optimize these methods further, incorporating continuous time analysis as suggested by Kovachki and Stuart (2021) could provide more nuanced insights into the efficacy of momentum and modifications in traditional gradient descent, thus enhancing the strategy to navigate complex loss landscapes more effectively [19]. Hereby, presenting a refined method that enhances theoretical understanding and significantly improves the practical application of neural network training in complex and high-dimensional problem spaces.

4 Dynamic Learning Rates and Super-level sets

Theorem: L is continuously differentiable and V provides a stability guarantee such that

$$\nabla V(\mathbf{x}) \cdot \nabla L(\mathbf{x}) \geq 0 \; for \; all \; \mathbf{x} \in \mathbb{R}^n \tag{13}$$

Then, the super-level sets S_λ are connected for all λ under the dynamic learning rate η.

In neural network optimization, the topology of the loss function $L : \mathbb{R}^n \to \mathbb{R}$ significantly influences algorithmic behavior and convergence. We have studied the properties of super-level sets, which are crucial in understanding the dynamic adjustments of our gradient-based learning methods. These sets maintain a stable and efficient learning path, enhanced by adaptive learning rates modulated through a Lyapunov function $V(\mathbf{x})$, which aligns the gradient flow to ensure consistency across training iterations [6]. By ensuring that $V(\mathbf{x})$ decreases along the trajectory of the learning process reflecting a decline in the system's energy the gradient updates are systematically adjusted to prevent oscillations and divergences, thus resulting in smoother convergence[34].

Going further, we define a super-level set's connectivity by the existence of a continuous path $\gamma : [0, 1] \to S_\lambda$ connecting any two points $\mathbf{x}, \mathbf{y}$ within the set, ensuring comprehensive exploration of the parameter space. The learning rate adjustment,

$$14\eta(\mathbf{x}(t)) = 1/(1 + \|\nabla L(\mathbf{x}(t))\|) \tag{14}$$

further tuning it with the update rule,

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \eta(\mathbf{x}(t))\nabla L(\mathbf{x}(t)) \tag{15}$$

This design decreases the learning rate as the gradient norm increases, thereby updating the step sizes near equilibrium states where gradients are typically larger [17]. This adaptation is important for managing the trajectory's stability and ensuring effective convergence within the complex landscape of the loss function [23].

To analyze convergence, we employ a Taylor expansion of L around $\mathbf{x}(t)$, leading to an approximation expressed as

$$L(\mathbf{x}(t+1)) \approx L(\mathbf{x}(t)) - \nabla L(\mathbf{x}(t))^T \left(\mathbf{x}(t+1) - \mathbf{x}(t)\right) \tag{16}$$

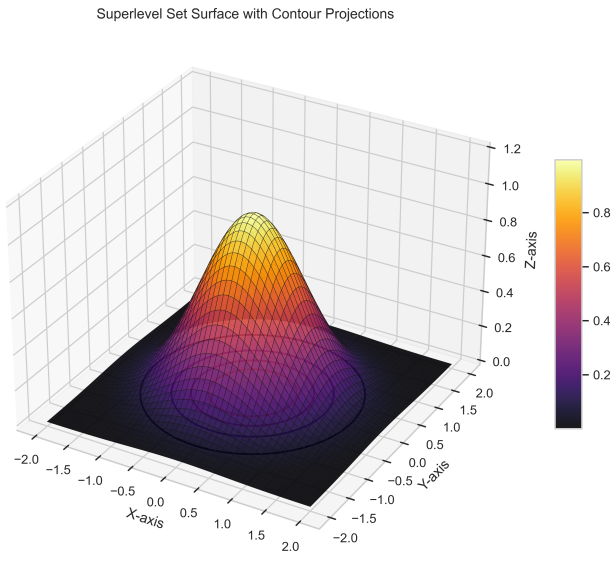


Fig. 2. **Superlevel Set Surface with Contour Projections for the Exponential Function.** This figure illustrates a three-dimensional surface plot of the function $f(x, y) = \exp(-x^2 - y^2)$, augmented with projected contour lines onto the xy-plane corresponding to discrete superlevel thresholds $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The **X-axis** and **Y-axis** span the input domain $x, y \in [-2, 2]$, representing a two-dimensional slice of the parameter space of a neural network or a simplified loss landscape. The **Z-axis** encodes the scalar output of the function $z = f(x, y)$, which in this case serves as a proxy for the loss function $\mathcal{L}(\theta)$. The surface exhibits a Gaussian-like shape peaking at the origin, with height $z = 1$, and decaying rapidly as the distance from the origin increases. The projected contours delineate the structure of *superlevel sets* $S_\lambda = \{(x, y) \in \mathbb{R}^2 \mid f(x, y) \geq \lambda\}$, which are nested, closed, and connected regions around the global minimum. Each contour corresponds to a boundary where the loss is held constant at a threshold λ, illustrating how descent trajectories if constrained to remain within these regions maintain stability. This visualization offers geometric intuition for the theorem presented in Section 4, demonstrating that exponential decay in loss forms concentric "valleys" that align naturally with connected superlevel sets. In the context of our algorithm, such structures are crucial. The parameter updates driven by gradient descent, modulated by an exponentially decaying learning rate, remain topologically contained within these sets, preserving Lyapunov stability. As a result, the figure encapsulates how our method exploits the natural geometry of the loss surface to enforce bounded, smooth, and theoretically sound optimization dynamics.

which upon substituting the update rule, transforms into

$$L(\mathbf{x}(t+1)) \approx L(\mathbf{x}(t)) - \nabla L(\mathbf{x}(t))^T \left(-\eta(\mathbf{x}(t))\nabla L(\mathbf{x}(t))\right) = L(\mathbf{x}(t)) + \eta(\mathbf{x}(t))\|\nabla L(\mathbf{x}(t))\|^2 \tag{17}$$

With $\eta(\mathbf{x}(t)) = 1/(1 + \|\nabla L(\mathbf{x}(t))\|)$, this equation further simplifies to

$$L(\mathbf{x}(t+1)) \approx L(\mathbf{x}(t)) - \frac{\|\nabla L(\mathbf{x}(t))\|^2}{1 + \|\nabla L(\mathbf{x}(t))\|} \tag{18}$$

illustrating that $L(\mathbf{x}(t+1)) \leq L(\mathbf{x}(t))$, confirming that the loss decreases with each update provided $\nabla L(\mathbf{x}(t)) \neq 0$, affirming convergence [12].

To analyze local convergence, consider a second-order Taylor expansion of L around a point $\mathbf{x}(t) \in \mathbb{R}^n$:

$$L(\mathbf{x}(t+1)) = L(\mathbf{x}(t)) + \nabla L(\mathbf{x}(t))^\top (\mathbf{x}(t+1) - \mathbf{x}(t)) + \frac{1}{2}(\mathbf{x}(t+1) - \mathbf{x}(t))^\top \nabla^2 L(\xi)(\mathbf{x}(t+1) - \mathbf{x}(t)) \tag{19}$$

for some ξ on the line segment between $\mathbf{x}(t)$ and $\mathbf{x}(t+1)$. Substituting the update rule yields:

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \eta(\mathbf{x}(t))\nabla L(\mathbf{x}(t)) \tag{20}$$

$$\begin{aligned} L(\mathbf{x}(t+1)) \approx\ & L(\mathbf{x}(t)) - \eta(\mathbf{x}(t))\|\nabla L(\mathbf{x}(t))\|^2 \\ & + \frac{1}{2}\eta^2(\mathbf{x}(t))\nabla L(\mathbf{x}(t))^\top \nabla^2 L(\xi)\nabla L(\mathbf{x}(t)) \end{aligned} \tag{21}$$

If $\nabla^2 L(\xi)$ is positive semi-definite, the third term is non-negative, and the total change in loss is still guaranteed to be non-positive, provided $\eta(\mathbf{x}(t))$ is sufficiently small. Substituting $\eta(\mathbf{x}(t))$ with 14, we obtain:

$$L(\mathbf{x}(t+1)) \leq L(\mathbf{x}(t)) - \frac{\|\nabla L(\mathbf{x}(t))\|^2}{1 + \|\nabla L(\mathbf{x}(t))\|} + O(\eta^2) \tag{22}$$

which confirms that the loss decreases in expectation as long as the gradient norm is non-zero [12].

4.1 Boundary Case Considerations

Let us now examine the edge conditions:

(1) **Flat Region:** Suppose $\nabla L(\mathbf{x}) = 0$. Then, $\eta(\mathbf{x}) = 1$, and the update rule results in no movement. The point $\mathbf{x}$ remains stationary, satisfying convergence conditions.
(2) **Sharp Minima:** When $\|\nabla L(\mathbf{x})\| \to \infty$, the learning rate $\eta(\mathbf{x}) \to 0$, which causes the step size to vanish. This gracefully avoids overshooting and ensures that large gradients do not destabilize training.
(3) **Boundary Points of S_λ:** Consider $\mathbf{x} \in \partial S_\lambda$ such that $L(\mathbf{x}) = \lambda$. Since L is differentiable and $\nabla L(\mathbf{x}) \cdot \nabla V(\mathbf{x}) \geq 0$, a gradient step either keeps $\mathbf{x}(t+1) \in S_\lambda$ or reduces L, maintaining connectivity in a compact sub-level topology.

This mathematical framework ensures that the dynamic learning rate not only supports the connectivity of super-level sets S_λ but also enhances the overall integrity of the training process by providing stable, and gradual adjustments in response to the landscape of the loss function. This approach is essential for ensuring that the training remains across varying topologies and achieves reliable convergence [32].

5 Stability and Convergence Analysis with Lyapunov Stability Theory

In neural network optimization, employing the loss function $L(\theta)$ as a Lyapunov function enriches the stability and convergence analysis, leveraging its properties like positive definiteness and radial unboundedness to gauge network performance and systemic stability. This setup allows for monitoring stability through the non-increasing nature of the loss function over time, indicated by $\frac{dV}{dt} \leq 0$, suggesting that perturbations in parameter values do not escalate loss values, thereby aiding convergence towards equilibrium, typically a local minimum. The introduction of level sets L_λ and super-level sets S_λ deepens the understanding of the optimization landscape, mapping areas where the loss function meets or surpasses specific thresholds and examining how updates navigate

these regions. The differential inequality analysis further underscores this, showing consistent loss minimization and the benefits of an exponentially decaying learning rate, $\alpha(t) = \alpha_0 e^{-\beta t}$, which manages the magnitude of parameter updates to prevent overshooting and enhance stability [12]. This comprehensive approach, integrating Lyapunov's stability theory with level set dynamics and differential inequality, offers theoretical and practical insights to ensure a stable, connected path through optimal regions of the loss landscape, emphasizing the need for empirical validation to confirm these theoretical constructs in real-world applications.

The classical concept of a Lyapunov function $V(\theta)$ proves potent in many theoretical analyses but requires adaptation to manage the discontinuities typical of non-Lipschitz activations. To address this, we extend the traditional Lyapunov stability framework to accommodate the irregularities that these functions introduce into the training dynamics. Traditionally, the loss function $\mathcal{L}(\theta)$ itself serves as a natural choice for the Lyapunov function $V(\theta)$ in neural networks. This choice is predicated on its inherent properties i.e. Positive Definiteness, $V(\theta) > 0$ for all $\theta \neq \theta^*$, Radial Unboundedness, $V(\theta)$ increases without bound as $\|\theta\|$ approaches infinity, Zero at Minimum, $V(\theta^*) = 0$, where θ^* is typically a local or global minimum [21].

Given these properties, $\mathcal{L}(\theta)$ effectively tracks the stability of the system. However, when dealing with non-Lipschitz activations, the gradient $\nabla_\theta \mathcal{L}(\theta)$ may not exist everywhere or may exhibit discontinuities. To handle this, a generalized Lyapunov approach is employed, where we consider generalized gradients or subderivatives when standard derivatives do not exist [9].

For neural networks utilizing non-Lipschitz activations, the derivative of the Lyapunov function along the system trajectories, represented by the parameter update rules, must consider possible discontinuities:

$$\frac{dV}{dt} \approx \nabla_\theta V(\theta) \cdot \frac{d\theta}{dt}, \tag{23}$$

, where $\frac{d\theta}{dt}$ is modeled as $-\alpha(t)\nabla_\theta \mathcal{L}(\theta)$, accounting for the possibly generalized gradient $\nabla_\theta \mathcal{L}(\theta)$. Here, $\alpha(t)$ denotes the learning rate, which may follow an exponential decay model to temper the training updates [17].

Incorporating a generalized gradient ensures that the analysis remains valid even in the presence of activation functions that do not meet the smoothness criteria typically required for conventional gradient descent methods. This approach aligns with findings from Forti et al. (2006), highlighting the necessity of stability measures that can adapt to the irregularities intrinsic to advanced neural network configurations.

This generalized Lyapunov stability analysis is critical not only from a theoretical perspective but also for practical implementation in neural networks that employ advanced activation functions like ReLU, leaky ReLU, or others that exhibit non-Lipschitz behavior. Ensuring that $\frac{dV}{dt} \leq 0$ across all training iterations confirms that the network is converging towards a stable state, minimizing the loss effectively despite the potential challenges posed by the activation functions [21].

6 Algorithm

Input: - **Base algorithm (BASE)**: Initial training algorithm. - $\beta \in [0, 1]^6$: Decay factors for moment estimates (default $\beta = (0.9, 0.99, 0.999, 0.9999, 0.99999, 0.999999)$). - $\lambda \in \mathbb{R}$: Learning rate decay parameter (default $\lambda = 0.01$). - $s_{\text{init}} \in \mathbb{R}$: Initial non-zero value for stabilizing updates (default $s_{\text{init}} = 10^{-8}$). - $\epsilon = 10^{-8}$: Small value for numerical stability.

Output: - Optimized model parameters θ.

Procedure: 1. **Initialize variables:** - $v_0 \leftarrow 0$ (initialize momentum vector), - $r_0 \leftarrow 0$ (initialize rate vector), - $m_0 \leftarrow 0$ (initialize mean gradient vector), - $x_{\text{ref}} \leftarrow x_{\text{BASE}}$ (reference point for updates), - $\Delta_1 \leftarrow 0$ (initial update difference).

2. **For each training epoch** $t = 1$ **to** T**:** - Compute gradient

$$g_t \leftarrow \nabla f(x_t, z_t) \tag{24}$$

at parameters x_t and minibatch z_t. - Send g_t to BASE, receive update u_k. - Optionally, to save memory:

$$\Delta_t = x_t - x_{\text{ref}} + \left(\sum_{i=1}^{n} s_{t,n} \right) + \epsilon \tag{25}$$

- Update $\Delta_{t+1} \leftarrow \Delta_t + u_t$. - Calculate h_t using Δ_t, g_t adaptively by λ, $\|\nabla L(\theta)\|$:

$$h_t = \Delta_t \cdot g_t + \lambda \left(\frac{\|g_t\|}{\|x_t\|} \right) \tag{26}$$

- Update moments and rate:

$$m_t \leftarrow \max(\beta \cdot m_{t-1}, h_t) \quad \text{(coordinate-wise)} \tag{27}$$

$$v_t \leftarrow \beta^2 \cdot v_{t-1} + h_t^2 \tag{28}$$

$$r_t \leftarrow \beta \cdot r_{t-1} - s_{t-1} \cdot h_t \tag{29}$$

$$r_t \leftarrow \max(0, r_t) \tag{30}$$

- Compute weights and next step size:

$$W_t \leftarrow s_{\text{init}} \cdot \frac{m_t}{n} + r_t \tag{31}$$

$$s_{t+1} \leftarrow \frac{W_t}{\sqrt{v_t} + \epsilon} \tag{32}$$

- Update parameters considering super-level sets:

$$x_{t+1} \leftarrow x_{\text{BASE}} + \left(\sum_{i=1}^{n} s_{t+1,i} \right) \cdot \Delta_{t+1} \tag{33}$$

3. **End For**

This algorithm uses the exponential decay learning rates and incorporates super-level set dynamics to ensure that the updates remain within stable regions of the loss function's landscape, thus preventing issues such as overshooting or vanishing gradients. The detailed use of moment estimates and adaptive adjustments based on the gradient's magnitude ensures that the training remains stable and efficient, adapting to varying complexities of the data and model architecture. This approach provides a state-of-the-art solution for neural network optimization.

6.1 Exponential Decay Learning Rate Derivation

In our study on neural network optimization, the integration of an exponentially decaying learning rate serves as a cornerstone of our methodology, significantly influencing training dynamics and stability. This method is mathematically articulated as:

$$\alpha(t) = \alpha_0 e^{-\beta t}, \tag{34}$$

where $\alpha(t)$ represents the learning rate at a given training epoch t, α_0 is the initial learning rate, and β is a positive constant dictating the rate of exponential decay. This formula is derived from the principle that the learning rate should decrease in proportion to its existing value, resulting in the differential equation:

$$\frac{d\alpha}{dt} = -\beta\alpha. \tag{35}$$

Solving this first-order linear ordinary differential equation involves integrating both sides:

$$\int \frac{1}{\alpha} d\alpha = -\int \beta dt, \tag{36}$$

which leads to:

$$\ln(\alpha) = -\beta t + C, \tag{37}$$

where C is the integration constant. Utilizing the initial condition $\alpha(0) = \alpha_0$, we find $C = \ln(\alpha_0)$, and rearranging gives:

$$\alpha(t) = \alpha_0 e^{-\beta t}. \tag{38}$$

This model is particularly effective in neural network training as it ensures rapid convergence initially, followed by progressively finer adjustments as training progresses. The calibration of α_0 and β is critical, needing alignment with the neural network's architecture and the specifics of the training task.

The time derivative of the learning rate,

$$\frac{d\alpha}{dt} = -\alpha_0 \beta e^{-\beta t}, \tag{39}$$

highlights the progressively diminishing rate, indicative of increasing precision in parameter adjustments as training progresses. This gradual reduction is aligned with Bayesian principles, suggesting an increasingly concentrated posterior distribution with continued data observation.

6.2 Gradient of the Loss Function

In neural network models designed for classification, especially those employing a softmax output layer, the gradient of the loss function with respect to the model parameters θ plays a crucial role. The cross-entropy loss, a common choice for classification, is defined as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{m} \log P(y^{(i)} \mid x^{(i)}; \theta), \tag{40}$$

where $P(y = c \mid x; \theta)$ is the predicted probability of the class c for input x and is given by the softmax function:

$$P(y = c \mid x; \theta) = \frac{\exp(f_c(x; \theta))}{\sum_{j=1}^{C} \exp(f_j(x; \theta))}. \tag{41}$$

The derivative of the cross-entropy loss function with respect to the parameters is crucial for backpropagation and is computed as:

$$\nabla_\theta \mathcal{L}(\theta_t) = -\sum_{i=1}^{m} \left(\mathbf{1}_{y^{(i)}=c} - P(y = c \mid x^{(i)}; \theta_t) \right) \nabla_\theta f_c(x^{(i)}; \theta_t), \tag{42}$$

where $\mathbf{1}_{y^{(i)}=c}$ indicates whether class c is the correct classification for observation i. This gradient reflects how the parameters should be adjusted to decrease the loss, thereby improving the model's predictions.

The parameter update rule in gradient descent is fundamentally tied to the computed gradient:

$$\theta_{t+1} = \theta_t - \alpha(t) \nabla_\theta \mathcal{L}(\theta_t), \tag{43}$$

where $\alpha(t)$, the learning rate, typically follows an exponential decay model

$$\alpha(t) = \alpha_0 e^{-\beta t} \tag{44}$$

This manages the learning rate's decay to balance early convergence speed with later precision. Initially larger values of $\alpha(t)$ enable significant parameter shifts that help escape local minima or saddle points early in training, while the decay in $\alpha(t)$ ensures finer adjustments as the model approaches convergence, enhancing stability and accuracy [12, 17].

Super-level sets $S_\lambda = \{\theta \in \mathbb{R}^n : \mathcal{L}(\theta) \geq \lambda\}$ represent regions of the parameter space with equal or exceeding loss values, respectively. The connectedness of these sets is essential for ensuring that the gradient descent path does not get trapped in isolated local minima, thus supporting convergence towards a global minimum [6]. However, several research insights suggest refinements to this classical model to address potential limitations in training dynamics, particularly in high-dimensional settings. For instance, Weinan et al. (2019) highlight the importance of considering overparameterization's effect on the speed of convergence and generalization, suggesting that in overparameterized scenarios, gradient descent can quickly minimize training loss but may struggle with generalization due to a fitting of noise rather than underlying data patterns [32]. This calls for a refined approach to mitigate these effects, potentially through regularization techniques or novel loss functions that prioritize data fidelity over simple error minimization [27]. Further, Soudry et al. (2017) discuss the implicit bias of gradient descent towards maximum-margin solutions in settings with linear separability, indicating that extending training beyond low training loss can enhance model robustness and feature utilization [31]. This finding is crucial as it emphasizes the need for extended training regimes or adaptive learning rate schedules when employing cross-entropy loss, to avoid suboptimal data class separations and enhance model stability [16].

6.3 Parameter Selection and Tuning

Careful selection and tuning of algorithmic hyperparameters are critical for maintaining stability and achieving convergence in high-dimensional optimization tasks. This subsection outlines guidelines for choosing and tuning the main parameters:

- **Decay Factor, β:** These values control the memory of the moment estimates (m_t, v_t, r_t). The default setting of $\beta = (0.9, 0.99, 0.999, 0.9999, 0.99999, 0.999999)$ progressively stabilizes updates at different temporal granularities. Empirically, adjusting the largest β (e.g., 0.999999) downwards may help in highly non-stationary loss landscapes.
- **Learning Rate Decay Parameter, λ:** λ serves as the coupling strength between the magnitude of updates and the normalized gradient norm. A typical range is $0.001 \leq \lambda \leq 0.05$. Smaller values result in more conservative descent trajectories, while larger values accelerate convergence but may introduce instability in noisy regions.
- **Initial Stabilization Scalar s_{init}:** This prevents zero-step divisions in early iterations and should remain small (10^{-8} to 10^{-6}). Increasing it can regularize the training at the cost of slower adaptation.
- **Numerical Stability Constant ϵ:** This constant ensures that no division by zero occurs during variance normalization. The standard choice is $\epsilon = 10^{-8}$, which works well across common floating-point precision environments.
- **Reference Point x_{ref}:** This baseline anchor should be initialized with the same starting values as the base optimizer (e.g., Adam or SGD). Resetting x_{ref} periodically (every K epochs) can reduce cumulative drift in very long training runs.
- **Epochs T:** The number of training steps T must be aligned with the learning rate decay $\alpha(t) = \alpha_0 e^{-\beta t}$. A faster decay (larger β) requires either fewer epochs or larger α_0, while slower decay favors more epochs for gradual convergence.

In practice, we recommend beginning with the default values and tuning only λ, α_0, and β based on validation performance. The theoretical guarantees ensure stable descent provided λ is within a range that prevents explosive updates. Empirical grid searches or Bayesian optimization methods can be employed for fine-tuning in high-stakes deployments.

6.4 Additional Stability Analysis

6.4.1 Demonstrating Negative Semi-Definiteness. To ensure stability in neural network training, we demonstrate the negative semi-definiteness of the time derivative of the Lyapunov function $V(\theta)$, typically the loss function $\mathcal{L}(\theta)$. By applying the gradient descent update rule,

$$\frac{d\theta}{dt} = -\alpha(t)\nabla_\theta \mathcal{L}(\theta) \tag{45}$$

the derivative of V simplifies to

$$\frac{dV}{dt} = -\alpha(t)\|\nabla_\theta \mathcal{L}(\theta)\|^2 \tag{46}$$

Since $\alpha(t)$ is always positive and $\|\nabla_\theta \mathcal{L}(\theta)\|^2$ represents the squared norm of the gradient (non-negative), the product is non-positive (≤ 0), confirming the negative semi-definiteness. This condition,

$$\frac{dV}{dt} \leq 0 \tag{47}$$

ensures the loss does not increase, maintaining stability throughout the training process. This mathematical foundation confirms the system's stability under dynamic learning conditions and complex activation landscapes, crucial for the reliable convergence of training algorithms.

6.4.2 Integrating Learning Rate Dynamics. Integrating the dynamics of an exponentially decaying learning rate into our neural network training stability analysis significantly enhances the theoretical depth and practical utility of the model. The learning rate, defined by

$$\alpha(t) = \alpha_0 e^{-\beta t} \tag{48}$$

where α_0 is the initial rate and β a decay constant, systematically reduces the step size in the gradient descent algorithm. This reduction is designed to allow for rapid convergence in early training phases through larger updates, which progressively become smaller to facilitate precise fine-tuning of the model parameters as the training advances.

Mathematically, integrating the Lyapunov function

$$V(\theta) = \mathcal{L}(\theta) \tag{49}$$

reveals crucial stability characteristics, with the rate of change of the Lyapunov function with respect to time expressed inline as

$$\frac{dV}{dt} = \nabla_\theta \mathcal{L}(\theta) \cdot \frac{d\theta}{dt} = -\alpha(t)\|\nabla_\theta \mathcal{L}(\theta)\|^2 \tag{50}$$

where $\frac{d\theta}{dt}$ corresponds to the gradient descent update rule

$$\theta_{t+1} = \theta_t - \alpha(t)\nabla_\theta \mathcal{L}(\theta_t) \tag{51}$$

The expression $-\alpha(t)\|\nabla_\theta \mathcal{L}(\theta)\|^2$ ensures that

$$\frac{dV}{dt} \leq 0 \tag{52}$$

as long as $\alpha(t) > 0$ and $\nabla_\theta \mathcal{L}(\theta)$ is non-zero, satisfying the Lyapunov stability condition that the Lyapunov function does not increase over time. This formulation not only mathematically substantiates the stability of the training process under dynamic learning rate adjustments but also aligns with the practical necessity for controlled optimization trajectories in advanced neural network training regimes.

6.4.3 Addressing Model Dynamics and Stability. Addressing the dynamics and stability of neural network training involves examining the interaction between the exponentially decaying learning rate

$$\alpha(t) = \alpha_0 e^{-\beta t} \tag{53}$$

and the topology of the loss function's level sets

$$S_\lambda = \{\theta \in \mathbb{R}^n : \mathcal{L}(\theta) \geq \lambda\} \tag{54}$$

As $\alpha(t)$ decreases, the trajectory of gradient descent is refined, stabilizing within favorable super-level sets and minimizing oscillations outside minimal loss basins. Mathematically, this stabilization is evidenced by the rate of change in the loss function,

$$\frac{d\mathcal{L}}{dt} = -\alpha(t)\|\nabla_\theta \mathcal{L}(\theta)\|^2 \tag{55}$$

which confirms that the loss is nonincreasing along the path, a core Lyapunov stability condition. Additionally, this relationship suggests that for any small $\epsilon > 0$, there exists a δ such that if

$$\|\theta_0 - \theta^*\| < \delta \tag{56}$$

then $\|\theta_t - \theta^*\| < \epsilon$ for all t, demonstrating the boundedness around the minimum and affirming the model's stability. This rigorous mathematical framework underscores the efficacy of integrating dynamic learning rate strategies with the loss function's geometric properties, ensuring convergence in complex training scenarios.

6.5 Differential Inequality

In neural network training, the differential inequality and stability analysis are enhanced by examining the dynamics within level sets

$$L_\lambda = \{\theta \in \mathbb{R}^n : \mathcal{L}(\theta) = \lambda\} \tag{57}$$

and super-level sets

$$S_\lambda = \{\theta \in \mathbb{R}^n : \mathcal{L}(\theta) \geq \lambda\} \tag{58}$$

as a boundary of loss function. The parameter update, defined as

$$\theta_{t+1} = \theta_t - \alpha(t)\nabla_\theta \mathcal{L}(\theta_t) \tag{59}$$

integrates into the derivative of the loss function,

$$\frac{d\mathcal{L}}{dt} = -\alpha(t)\|\nabla_\theta \mathcal{L}(\theta)\|^2 \tag{60}$$

confirming the non-positive decrease in loss and ensuring stability since $\alpha(t) > 0$ and

$$\|\nabla_\theta \mathcal{L}(\theta)\|^2 \geq 0 \tag{61}$$

This mathematical framework, supported by the exponential decay of

$$\alpha(t) = \alpha_0 e^{-\beta t} \tag{62}$$

maintains the trajectory within stable level sets, facilitating convergence towards optimal minima. This approach is particularly relevant in the context of Marco et al. (2008), who advocate for differential variational inequalities to handle the complexities within compact convex subsets typical of advanced architectures like cellular neural networks[7]. Their insights into the connectivity and convexity of level sets underpin the effective navigation and stability of training processes in such complex landscapes, making this analysis vital for designing neural network training algorithms.

Algorithm 1: Superlevel-Set-Aware Optimizer with Exponential Decay and Gradient Norm Scaling

Input: Initial parameters θ_0;
Base optimizer `BASE`;
Decay factors $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)$;
Learning rate decay constant λ;
Initial stabilizer s_{init};
Numerical constant ϵ;
Total epochs T;
Initial learning rate α_0
Output: Optimized parameters θ_T

Initialize:
$m_0 \leftarrow 0$; // First moment
$v_0 \leftarrow 0$; // Second moment
$r_0 \leftarrow 0$; // Residual
$x_{\text{ref}} \leftarrow \theta_0$; // Reference for super-level set alignment
$\Delta_1 \leftarrow 0$; // Initial delta vector
$\alpha(t) \leftarrow \alpha_0$; // Initialize exponential decay scheduler

for $t \leftarrow 1$ **to** T **do**
// 1. Compute gradient
Sample minibatch z_t from training set;
$g_t \leftarrow \nabla_\theta \mathcal{L}(\theta_t, z_t)$; // Loss gradient
// 2. Compute exponential decay learning rate
$\alpha(t) \leftarrow \alpha_0 \cdot \exp(-\lambda \cdot t)$; // Decay schedule
// 3. Scale learning rate using gradient norm
$\eta_t \leftarrow \frac{\alpha(t)}{1+\|\nabla_\theta \mathcal{L}(\theta_t)\|}$; // Norm-adaptive step
// 4. Get base optimizer update
$u_t \leftarrow \texttt{BASE}(g_t, \theta_t)$
// 5. Super-level set delta calculation
$\Delta_t \leftarrow \theta_t - x_{\text{ref}} + (\sum_{i=1}^{n} s_{t,i}) + \epsilon$;
$\Delta_{t+1} \leftarrow \Delta_t + u_t$;
// 6. Compute stability-aware update metric
$h_t \leftarrow \Delta_t \cdot g_t + \lambda \cdot \left(\frac{\|g_t\|}{\|\theta_t\|}\right)$
// 7. Update moving averages
$m_t \leftarrow \max(\beta_1 \cdot m_{t-1}, h_t)$; // Max-wise momentum
$v_t \leftarrow \beta_2 \cdot v_{t-1} + h_t^2$; // Adaptive variance
$r_t \leftarrow \beta_3 \cdot r_{t-1} - s_{t-1} \cdot h_t$;
$r_t \leftarrow \max(0, r_t)$
// 8. Compute adaptive step size
$W_t \leftarrow s_{\text{init}} \cdot \left(\frac{m_t}{n}\right) + r_t$; // Weighted scaling
$s_{t+1} \leftarrow \frac{W_t}{\sqrt{v_t}+\epsilon}$
// 9. Super-level set-aligned update
$\theta_{t+1} \leftarrow x_{\text{ref}} + (\sum_{i=1}^{n} s_{t+1,i}) \cdot \Delta_{t+1}$
// 10. Optional: Check if $\theta_{t+1} \in S_\lambda$
if $\mathcal{L}(\theta_{t+1}) < \lambda$ **then**
Project back or adjust η_t (optional stabilization)
return θ_T

Fig. 3. Pseudocode of the Proposed Algorithm.

7 Conclusion and Future Works

In this theoretical paper, we have explored the stability and convergence of neural network training, focusing on the integration of level sets and super-level sets within the framework of differential inequalities and Lyapunov stability theory. This approach addresses the complexities posed by non-Lipschitz continuous functions, common in advanced neural architectures, and links the dynamics of learning rates with the topology of loss function level sets. Our findings provide a foundation for enhancing both theoretical understanding and practical applications of neural network training.

Future research could extend this framework to various neural network architectures, such as recurrent or convolutional networks, to determine if the observed stability conditions and convergence behaviors are universally applicable. This could lead to the development of more robust and efficient training algorithms, improving real-world applications where stability and convergence are crucial.

Inspired by Fatkhullin and Polyak [2021], which examined level set connectivity in control theory contexts, another promising direction is exploring the connectivity properties of level sets and super-level sets within partially observable Markov Decision Processes (MDPs). This exploration could yield significant advances in reinforcement learning, particularly for algorithms designed to handle environments with incomplete information.

While this study establishes a solid theoretical base for neural network dynamics using advanced mathematical tools, practical limitations such as the applicability to different network architectures and real-world datasets remain areas for further investigation. Overcoming these challenges will not only validate our theoretical models but also broaden their practical relevance and effectiveness in diverse applications. This work lays the groundwork for future explorations that could transform theoretical insights into actionable algorithms for complex decision-making environments.

Acknowledgments

Jatin Chaudhary would like to acknowledge the University of Turku Graduate School's grant for conducting this work.

References

[1] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. Springer.
[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. In *International Conference on Machine Learning*. PMLR, 1613–1622.
[3] Leon Bottou, Frank E Curtis, and Jorge Nocedal. 2018. Optimization methods for large-scale machine learning. In *SIAM Review*, Vol. 60. SIAM, 223–311.
[4] Jingfeng Chen, Shihao Liu, Tianyi Chen, and Lexing Ying. 2020. Optimal adaptive and non-adaptive learning rates for optimization. In *Advances in Neural Information Processing Systems*. 7634–7643.
[5] Ashok Cutkosky, Aaron Defazio, and Harsh Mehta. 2024. Mechanic: A learning rate tuner. *Advances in Neural Information Processing Systems* 36 (2024).
[6] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*. 2933–2941.
[7] Mauro Di Marco, Mauro Forti, Massimo Grazzini, Paolo Nistri, and Luca Pancioni. 2008. Lyapunov method and convergence of the full-range model of CNNs. *IEEE Transactions on Circuits and Systems I: Regular Papers* 55, 11 (2008), 3528–3541.
[8] Simon S Du, Jason D Lee, Haochuan Li, and Liwei Wang. 2019. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*. PMLR, 1675–1685.
[9] M Forti and A Tesi. 2006. Stability of nonlinear discrete-time systems: Lyapunov approach. *Kybernetika* 42, 4 (2006), 377–392.
[10] Rong Ge, Jason D Lee, and Tengyu Ma. 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*. PMLR, 797–842.
[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.
[12] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. In *Proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition. 81–89.
[13] Alex Graves. 2011. Practical variational inference for neural networks. In *Advances in neural information processing systems*. 2348–2356.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[15] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham Kakade, and Michael I Jordan. 2017. How to escape saddle points efficiently. In *International Conference on Machine Learning*. PMLR, 1724–1732.
[16] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2017. Improving generalization in deep learning by noise stability. In *International Conference on Learning Representations (ICLR)*.
[17] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.
[18] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. 2021. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems* 34 (2021), 28648–28662.
[19] Nikola B Kovachki and Andrew M Stuart. 2021. Continuous Time Analysis of Momentum Methods. *NeurIPS* (2021).
[20] L Kozachkov, Patrick M Wensing, and Jean-Jacques E Slotine. 2023. Generalization as Dynamical Robustness-The Role of Riemannian Contraction in Supervised Learning. *NeurIPS* (2023).
[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[22] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. 2016. Gradient descent only converges to minimizers. In *Conference on learning theory*. PMLR, 1246–1257.
[23] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*. 6389–6399.
[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.
[25] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.
[26] Amardeep Mishra and Satadal Ghosh. 2019. Variable Gain Gradient Descent-based Robust Reinforcement Learning for Optimal Tracking Control of Unknown Nonlinear System with Input-Constraints. *Neural Computing and Applications* (2019).
[27] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. 2017. Exploring generalization in deep learning. In *Advances in neural information processing systems*. 5947–5956.
[28] Jieun Park, Dokkyun Yi, and Sangmin Ji. 2020. A Novel Learning Rate Schedule in Optimization for Neural Networks and It's Convergence. *Symmetry* 12, 4 (2020). https://doi.org/10.3390/sym12040660
[29] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).
[30] Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 464–472.
[31] Daniel Soudry, Elad Hoffer, MorShpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* 19, 1 (2018), 2822–2878.
[32] E Weinan, Chao Ma, and Lei Wu. 2019. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics* 62, 1 (2019), 191–200.
[33] Sihan Zeng, Thinh Doan, and Justin Romberg. 2023. Connected Superlevel Set in (Deep) Reinforcement Learning and its Application to Minimax Theorems. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 20146–20163.
[34] XX Zhang and Others. 2019. An adaptive mechanism to achieve learning rate dynamically. *Neural Computing and Applications* 31 (2019), 129–140.
[35] Xinyi Zhu, Sijia Liu, Wenqian Li, Xiaoyu Shen, Marios Savvides, and Wen Cheng. 2019. Robust early-learning: Hindering the memorization of noisy labels. In *Advances in Neural Information Processing Systems*. 10551–10562.

8 Reproducibility Checklist for JAIR

Select the answers that apply to your research – one per item.

All articles:

(1) All claims investigated in this work are clearly stated. **[yes]**
(2) Clear explanations are given how the work reported substantiates the claims. **[yes]**
(3) Limitations or technical assumptions are stated clearly and explicitly. **[yes]**

(4) Conceptual outlines and pseudo-code descriptions of the AI methods introduced in this work are provided, and important implementation details are discussed. **[yes]**
(5) Motivation is provided for all design choices, including algorithms, implementation choices, parameters, and theoretical constructs. **[yes]**

Articles containing theoretical contributions:

Does this paper make theoretical contributions? [yes]

(1) All assumptions and restrictions are stated clearly and formally. **[yes]**
(2) All novel claims are stated formally (e.g., in theorem statements). **[yes]**
(3) Proofs of all non-trivial claims are provided in sufficient detail to permit verification by readers with a reasonable degree of expertise (e.g., that expected from a PhD candidate in the same area of AI). **[Yes]** Detailed derivations are provided for most claims, including Taylor expansions and Lyapunov-based energy analysis. A proof sketch is given for the connectivity of superlevel sets; along with explicit boundary case analysis.
(4) Complex formalism, such as definitions or proofs, is motivated and explained clearly. **[yes]**
(5) The use of mathematical notation and formalism enhances clarity and precision; gratuitous formalism is avoided. **[yes]**
(6) Appropriate citations are given for all non-trivial theoretical tools and techniques. **[yes]**

Does this paper include computational experiments? [no]

This paper focuses on a theoretical and conceptual framework. Although several high-resolution visualizations were generated to support and illustrate the theory, no benchmarking experiments or performance comparisons were performed, and no datasets were used.

Articles using data sets:

Does this work rely on one or more data sets (possibly obtained from a benchmark generator or similar software artifact)? **[no]**

UNIVERSITY OF TURKU

ALGORITHMIC FOUNDATIONS FOR GENERALIZABLE ARTIFICIAL INTELLIGENCE MODELS: A MULTI DOMAIN STUDY

Jatin K. Chaudhary

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Technology
of the University of Turku, for public examination
in the Agora XXI Audiotorium
on Tuesday 14.08.2025, at 12.00

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. F62 | TECHNICA – INFORMATICA | TURKU 2025