SMART DENOISING FOR RECURRENT NEURAL NETWORK OPTIMIZATION

Jakob Jelenčič

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Dunja Mladenić, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Ljupčo Todorovski, Chair, Faculty of Mathematics and Physics, University of Ljubljana; Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Jasminka Dobša, Member, Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
Prof. Dr. Nada Lavrač, Member, Jožef Stefan Institute, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Jakob Jelenčič

SMART DENOISING FOR RECURRENT NEURAL NETWORK OPTIMIZATION

Doctoral Dissertation

PAMETNO RAZŠUMLJANJE PRI OPTIMIZACIJI POVRATNIH NEVRONSKIH MREŽ

Doktorska disertacija

Supervisor: Prof. Dr. Dunja Mladenić

Ljubljana, Slovenia, June 2024

To BiBi

Acknowledgments

I would like to express my gratitude to all those who provided me the possibility to complete this thesis. A special word of thanks to my supervisor, Dunja, whose guidance was invaluable throughout the course of this research. Without her guidance, knowledge and most importantly patience none of this would be possible. Thank you for all committee members and extra gratitude to Ljupčo for extra help and guidance. I would like to thank my colleagues at Department for Artificial Intelligence and Slovenian National founding for young researchers.

Of course, special thank you goes without question to Ema, for emotionally supporting me, and my family for always having my back. Especially to you dad, always backing me, supporting every long shot I came up with.

Abstract

This thesis introduces a new optimization method based on deep learning, designed for data influenced by random processes. The main contribution of this method is the combination of advanced noise reduction techniques with recurrent neural network models, which helps to prevent the common problem of overfitting seen when there is not much training data available. This approach greatly improves how the model optimizes data representation and target outcomes, and enhances the model's performance by managing noise effectively.

The success of this method is thoroughly tested on two different types of data, financial and textual, demonstrating its flexibility and effectiveness. For financial data, it is applied to a large dataset containing equity prices over twenty years, which forms a large complex process where each equity follows unique random process. For textual data, the method is used on a dataset about required job skills from different languages spanning over five years, to show how it can be applied in more predictable situations.

This thesis makes important contributions to the field of deep learning in particular in recurrent neural networks, by creating a new method that effectively uses a sophisticated approach to noise management in neural networks. This method not only deals with the challenges of modeling random processes but also opens up new possibilities for making accurate predictions across different types of data.

The evaluation of a proposed approach includes three main parts: a formal definition where each part of the model is defined in detail, a statistical analysis using the Wilcoxon page rank test to show that our method performs better than existing models, and an in-depth insight into how the model's parameters can be adjusted and visualized. This detailed assessment shows that our approach performs better than current models and significantly improves the use of recurrent neural networks in handling random processes.

Povzetek

Doktorska disertacija predstavi novo optimizacijsko metodo, ki temelji na globokem učenju in je oblikovana za podatkovne zbirke z visoko stopnjo šuma. Glavni znanstveni prispevek je kombinacija pametnega odstranjevanja šuma za povratne nevronske mreže, ki preprečuje težave s prekomernim prilagajanjem podatkom, ko ni na voljo dovolj učnih podatkov. Metoda prispeva k boljši predstavitvi podatkov in napovedovanju ciljne spremenljivke.

Uspešnost metode je ovrednotena na dveh vrstah podatkov, in sicer finančnih in tekstovnih. Finančni nabor podatkov je sestavljen iz več kot dveh tisoč delnic, ki zajemajo več kot dvajset let zgodovine gibanja finančnih trgov, pri čemer vsak nabor podatkov predstavlja unikaten stohastični proces. Pri tekstovnih podatkih je metoda ocenjena na oglasih za delovna mesta, ki izvirajo iz šestnajstih različnih držav in več jezikov. Tekstovni podatki vključujejo več kot pet let zgodovine.

Ocenjevanje prispekov disertacije je razdeljeno na tri dele: formalne definicije, statistično analizo in vizualizacijo ter raziskovanje prostora parametrov. S tem se potrdi prispevek metode k znanosti. Disertacija pomembno prispeva k področju globokega učenja, zlasti na področju povratnih nevronskih mrež, z oblikovanjem nove metode, ki učinkovito uporablja prefinjen pristop za obvladovanje šuma v nevronskih mrežah. Metoda se ne spopada le z izzivi modeliranja stohastičnih procesov, ampak tudi odpira nove možnosti za natančne napovedi pri različnih vrstah podatkov.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

AI	...	artificial intelligence
RNN	...	recurrent neural network
LSTM	..	Long-Short term memory
ts	...	Time series
np	...	Number of parameters
bs	...	Batch size
chol	...	Cholesky decomposition, a mathematical function for matrix operations
matmul	.	Matrix multiplication operation
mvn	...	Multivariate normal distribution
sd	...	Standard deviation
W_{ae}	...	Weighting factor for autoencoder component in an optimization algorithm
GLM	...	Generalized Linear Model, a flexible generalization of ordinary linear regression
GPU	...	Graphics Processing Unit, commonly used to accelerate deep learning tasks
Adam	...	A method for stochastic optimization, used for training deep learning models
PCR	...	Premium Capture Rate, a metric used to assess the effectiveness of options trading strategies in capturing premiums
MDD	...	Maximum Drawdown, represents the maximum peak-to-trough decline in the value of a trading account or portfolio during a specific period
MAR	...	MAR Ratio, a critical risk-adjusted performance metric used to evaluate the profitability of an investment strategy relative to its risk exposure
SPX	...	Standard & Poor's 500 Index, an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ
DTE	...	Days to Expiration, refers to the number of days until an options contract expires

Chapter 1

Introduction

Noise is an inherent part of stochastic processes, which significantly complicates the analysis of data in real-world use cases. As the trend towards employing machine learning and especially deep learning intensifies, the challenge of effectively handling this noise becomes more critical. This thesis focuses on developing robust technique that can filter out noise and reveal the underlying patterns crucial for accurate predictions. Traditional statistical methods often fall short in managing the complexity and variability of noisy data, leading to increased interest in advanced deep learning approaches. These approaches aim to enhance model reliability and decision-making by improving the reliability of data representation and reducing the risk of overfitting.

Stochastic processes, characterized by their inherent randomness and unpredictability, are fundamental to numerous fields such as finance, meteorology, telecommunications, and biological systems. Accurate modeling of these processes is crucial for reliable predictions and informed decision-making. Traditional statistical methods often struggle to capture the complexity and variability inherent in stochastic data, particularly when historical data is limited. In recent years, deep learning has emerged as a powerful tool for modeling complex, non-linear relationships within data. However, a significant challenge in applying deep learning to stochastic processes is the issue of noise in the data. Noise can obscure underlying patterns, leading to models that overfit to the noise rather than generalizing to new data. Overfitting compromises the model's predictive performance and limits its practical applicability.

Denoising techniques have gained prominence as a solution to this challenge. Denoising involves filtering out noise from data, thus enhancing the quality and reliability of the input signals. By improving the quality of the data, denoising helps make the underlying patterns more discernible to deep learning models. This, in turn, reduces the risk of overfitting and enhances the model's ability to generalize from training data to unseen scenarios. Various methods, including autoencoders and convolutional neural networks, have been employed to effectively denoise data, each offering unique advantages in handling different types of noise.

Our research focuses on optimizing these denoising techniques within deep learning frameworks to improve the robustness and accuracy of models designed for stochastic processes. By rigorously testing these methods on diverse real-world datasets, we aim to demonstrate their effectiveness in achieving superior predictive accuracy and reliability. This advancement not only contributes to the field of deep learning but also opens new avenues for efficient and accurate predictions in applications dealing with stochastic data.

In the financial world, stochastic modeling plays a crucial role in predicting the behavior of complex systems characterized by uncertainty and randomness. Financial markets, in particular, exhibit intricate and dynamic relationships between various factors, making traditional deterministic models inadequate. Stochastic modeling offers a powerful approach to capture the inherent uncertainties and non-linearities in financial data. One of the primary challenges in stochastic modeling for financial data is accurately predicting short-term trends. Financial markets are influenced by a multitude of factors, including economic indicators, geopolitical events, investor sentiment, and unforeseen shocks, leading to highly volatile and unpredictable behavior.

Recently, there has been a surge in interest in employing deep learning techniques, such as recurrent neural networks (RNNs) [1] and transformers [2], to tackle the complexities of financial time series prediction. RNNs, especially the Long Short-Term Memory (LSTM) variant [1], have demonstrated promising results in capturing temporal dependencies in sequential data. Transformers, initially developed for natural language processing tasks [2], are now being explored for their potential in financial time series forecasting.

Despite the potential benefits of deep learning models, they are susceptible to overfitting, a common problem in situations with limited data and complex models. Overfitting occurs when the model memorizes noise and idiosyncrasies in the training data, resulting in poor generalization to unseen data. In financial forecasting, the ability to generalize accurately to new data is of utmost importance for making informed investment decisions.

To address the challenge of overfitting in deep learning models for financial time series prediction, researchers have explored various techniques. One effective approach involves the addition of noise during the training process [3]. Noise addition serves as a regularizing mechanism that prevents the model from fitting the training data too closely and encourages it to capture more meaningful patterns and relationships. By injecting controlled levels of noise into the input data or the model's hidden states, the model becomes less sensitive to minor fluctuations and outliers present in the training data, thereby enhancing its ability to generalize to unseen data. This regularization technique helps strike a balance between fitting noise and capturing true underlying trends, ultimately leading to more accurate and robust predictions. Moreover, by adapting the amount and type of noise added, practitioners can control the trade-off between model complexity and generalization performance, tailoring the approach to the specific characteristics of the financial data at hand.

1.1 Motivation

Stochastic processes refer to statistical phenomena that comprise random variables ordered based on a given index, usually time. These processes evolve over time following probabilistic laws, with non-deterministic time series reflecting their manifestation. Stochastic modeling introduces randomness into the models, such as initializing weights, selecting features, or introducing random input perturbations to comprehend their effect on outcome probability. Although stochastic modeling inherently introduces randomness into models, a probability distribution of the outputs can be obtained by repeating the process multiple times. This probability distribution serves as a basis for realistic future scenario assessments, which facilitate meaningful decision-making. Common use cases include forecasting stock prices and predicting cryptocurrency values. Nonetheless, accurately predicting future outcomes while considering the inherent randomness of a given context remains a challenge.

1.1.1 Challenges

Overfitting poses a common challenge when developing machine learning models, with models learning to make good predictions on training data but losing their ability to generalize to a test set. To circumvent this challenge, adding correlated noise to the training data can help the algorithm learn the information signal while ignoring white noise. In this regard, this study investigates adding correlated noise and latent representation to the optimization process. Our research shows that while optimizing an algorithm with a latent representation does not always lead to better results, adding correlated noise during the training process yields clear benefits. This technique is applicable to any recurrent deep learning architecture.

1.2 Overview of Related Work

The field of probability theory and stochastic processes has witnessed remarkable growth since 1950's and has become increasingly relevant in a variety of fields. One such application is in the valuation of financial derivatives using the Black-Scholes model, which is a well-known model based on stochastic processes [4]. Despite the popularity of the Black-Scholes model, it is not without its limitations, such as the volatility smile [5]. As a result, it is still an open problem to accurately predict the short-term trend of a stochastic process. It is worth noting that modeling stochastic processes alone may not be a perfect solution, as observed by some researchers [6]. Assessing the current state of the art of applied stochastic processes in the financial and insurance industry is always a challenge due to the highly competitive nature of the industry. Published research often lags behind what is actually being used in practice. However, a study by Madan et al. [7] provides a comprehensive overview of the field, focusing on Brownian motion and discussing Markov processes such as local volatility and local Lévy processes, as well as extensions to Lévy and Sato processes. While stochastic processes have become an indispensable tool in various domains, including finance and insurance, there is still much room for improvement in terms of their predictive power and practical application. Researchers and practitioners continue to develop new models and methods to better understand and utilize stochastic processes in real-world scenarios.

Time series prediction and classification have always been complex challenges in the field of data science and extensively researched problems [8]. Time series forecasting is important for many applications, including economic forecasting, stock market prediction, weather forecasting, and many more. However, it is a challenging task, as the underlying data-generating processes may be complex and non-linear, and there may be various external factors that can influence the time series. There are various methods available to predict the next step of a time series, ranging from traditional statistical models to deep learning techniques. Autoregressive Integrated Moving Average (ARIMA) [9] is a well-known statistical method that has been widely used for time series analysis and forecasting [10]. ARIMA is a derivative of ARMA (Autoregressive Moving Average), which combines two models - the autoregressive (AR) model and the moving average (MA) model - to capture the linear dependencies and the time-varying nature of the data. ARIMA adds an integration component to ARMA, which makes it more flexible to handle non-stationary time series.

With the recent advances in deep learning, many researchers have explored the use of recurrent neural networks (RNNs) for time series prediction [1]. RNNs are a type of neural network that can capture the temporal dependencies of the data by using feedback connections. Long Short-Term Memory (LSTM) is a popular variant of RNNs that can handle long-term dependencies and mitigate the vanishing gradient problem. LSTMs have been shown to be effective in many time series prediction tasks, including speech recognition, language translation, and financial forecasting. Recurrent structures decrease in popularity with the advent of attention structure. Transformers are a type of deep learning model that has gained popularity in natural language processing (NLP) due to their ability to handle long sequences of text and capture the contextual relationships between words [2]. While Transformers are superior in natural language processing, they may not be the best solution for all time series forecasting problems when data is limited.

Deep learning models can be prone to overfitting, especially when the data is limited, and the model is complex. Overfitting occurs when the model fits the training data too closely, and it fails to generalize to new data. In the case of time series forecasting, overfitting can be a significant problem, as the goal is to make accurate predictions for unseen data. To overcome the limitations of deep learning models, researchers have explored combining traditional statistical models with neural networks to create hybrid models. One such model is TM-ARMA, which combines ARIMA and neural networks to enhance prediction accuracy [11]. TM-ARMA utilizes the ARIMA model to capture the linear elements of the time series, while the neural network component handles the non-linear components. This approach has shown promising results in predicting the Chinese equities market during the Covid shock [11]. Another study has explored the use of Convolutional Neural Networks (CNNs) with 2-D histograms generated from quantified datasets to predict stock movements in the Indian stock market [12]. This approach has shown to be effective in predicting stock prices and outperforms traditional time series models. Finally, there is research done on the use of transformers in financial domain, Cheng et al. [13] showed their potential in research forecasting quantiles in multi-objective framework.

1.3 Hypotheses and Aims

1.3.1 Aims

Develop a Novel Methodology for Neural Network Optimization Based on Smart Denoising

The primary objective of this dissertation is to develop an innovative methodology for optimizing neural networks, which incorporates a smart noise structure into the optimization process. This approach is specifically crafted to tackle the intricacies and challenges posed by stochastic modeling. By integrating advanced denoising techniques within the neural network framework, the methodology seeks to enhance the network's ability to process and learn from data characterized by inherent randomness and noise variability.

Empirical analysis focused on financial data

The next objective is to conduct thorough empirical evaluations to assess the performance of the proposed methodology against current leading techniques in stochastic modeling. The initial focus will be on the financial sector, specifically on the prediction of stock movements. By rigorously analyzing the performance metrics, the study aims to demonstrate the effectivness and strengths of the novel methodology in real-world financial environments.

Assess Generalizability to Text Datasets

Finally, the aim is to assess the generalizability and versatility of the proposed methodology beyond its initial financial applications. The research will extend the novel denoising approach to multiple skill demand datasets, demonstrating the predictive power on less stochastic type of datasets.

1.3.2 Hypotheses

Hypothesis 1.

The smart denoising optimization exhibits superior performance compared to the current state-of-the-art methods in the context of stochastic processes when modeled by recurrent neural networks in financial data.

Hypothesis 2.

The proposed method demonstrates superior performance when applied to non-stochastic text data, surpassing the capabilities of existing state-of-the-art approaches.

1.4 Scientific Contributions

The scientific contributions of this thesis is in the area of modeling recurrent neural networks with denoising techniques. We provide evaluations on two types of data to illustrate the versatility and effectiveness of these models. Specifically, our analysis includes a detailed examination of financial data to showcase the applicability of the proposed approach to stochastic datasets, as well as an exploration of text data, using a skill demand dataset as a more deterministic example. In the following sections, we will offer a brief discussion and evaluation of each application, demonstrating the robust capabilities of our proposed modeling techniques in different contexts.

The main contributions of this thesis are listed below.

SC1 - Formal definition and development of the proposed methodology

Formal definition and development of a novel optimization technique based on smart denoising specialized for recurrent neural networks.

SC2 - Evaluation of the proposed methodology on financial data

Evaluation of the proposed methodology using real-world financial datasets. These datasets represent a complex, unknown stochastic process and include data from thousands of equity prices spanning over a twenty year period.

SC3 - Evaluation of the proposed methodology on text data

Evaluation of the proposed methodology using real-world example of text data, specifically skill demand datasets derived from multilingual job postings over sixteen countries spanning over five year period.

1.5 Thesis Structure

Thesis begins with an introduction, motivation and overview of the aims, hypotheses and scientific contributions.

Chapter 2 introduces the proposed methodology, addressing SC1. This chapter details each component of the proposed method, establishing the framework for its application and evaluation.

Chapter 3 focuses on the evaluation methodology. It outlines the structure of the model and details an ablation study. Additionally, it presents five hypotheses aimed at evaluating the method across various datasets to demonstrate SC2 and SC3. This chapter also discusses the analytical methods employed and describes the visualization techniques for parameter space, which are crucial for understanding the behavior of the proposed method from different perspectives.

Chapter 4 empirically evaluates SC2 and SC3. The chapter begins with a description of the datasets used, including financial and textual datasets focusing on skill demand. It then proceeds to present results from the ablation study and continues with the evaluation of the five hypotheses. The analysis of the parameter space is conducted first on financial data, followed by textual data.

Chapter 5 transitions from statistical analysis to practical applications in real-world scenarios. This chapter is divided into two sections: the first section demonstrates enhancements to a profitable, state-of-the-art trading strategy. The second section explores the application of the methodology to real-world AI demand analysis modeled by different countries, illustrating the practical benefits of the research, thus demonstrating the strength of SC2 and SC3 through practical applications.

Chapter 6 concludes the thesis by reflecting on the main scientific contributions. It summarizes the objectives achieved and provides directions for future research, emphasizing the implications and potential extensions of the work.

Chapter 2

Proposed Methodology for Smart Denoising Optimization

In this section, we outline and present the contributions defined in SC1. Our initial formulation of the proposed method was introduced in the journal paper "Improving modeling of stochastic processes by smart denoising published" [14] in Informatica journal and further discussed at the SiKDD conference paper "Modeling stochastic processes by simultaneous optimization of latent representation and target variable" [15]. Comprehensive analyses validating the effectiveness of our approach have been documented in the journal paper published in Information Sciences (Impact Factor: 8.1, CiteScore: 13.4, 2024) titled "Improving stochastic models by smart denoising and latent representation optimization" [16]. In the Informatica and SiKDD paper we introduced the method and showed the initial results, while in the paper published in Information Sciences journal we presented extensive parameter analysis and a statistical proof how the method improves the predictive power based on the type of data being modeled.

In the paper titled "Modeling stochastic processes by simultaneous optimization of latent representation and target variable" [15] we delineated the proposed method and demonstrated its efficacy on the equity and cryptocurrency datasets. We conducted a comparative analysis of the proposed method against established models including Random Forest, Generalized Linear Models (GLM), and non-noisy LSTM, illustrating the superior performance of our approach. Additionally, we explored an unsupervised example to underscore the significance of noise correlation as a valuable enhancement and discussed strategies for optimizing model fit to enhance performance on unseen test data.

In the paper "Improving Stochastic Models by Smart Denoising and Latent Representation Optimization," we present a more detailed analysis than that found in the Informatica study. We rigorously tested our approach using five varied real-world datasets, structuring the study into three parts: an ablation study to verify each component of our method, statistical analysis employing the Wilcoxon page rank test to demonstrate the superiority of our method over existing approaches, and an in-depth exploration of parameter visualization and fine-tuning. Our comprehensive evaluation shows that our method not only outperforms current techniques but also significantly improves the effectiveness of deep learning models in analyzing stochastic processes. This research underscores the robustness of our approach and its capacity to overcome existing challenges in modeling stochastic processes, leading to more precise and efficient predictive models.

The main motivation for the proposed method is in controlling the overfit of data. Overfitting poses a significant challenge in machine learning, with models frequently capturing the peculiarities of the training data rather than the underlying generalizable patterns. To address this issue, we introduce a novel three-component methodology designed to mitigate overfitting while boosting the model's performance on new, unseen data.

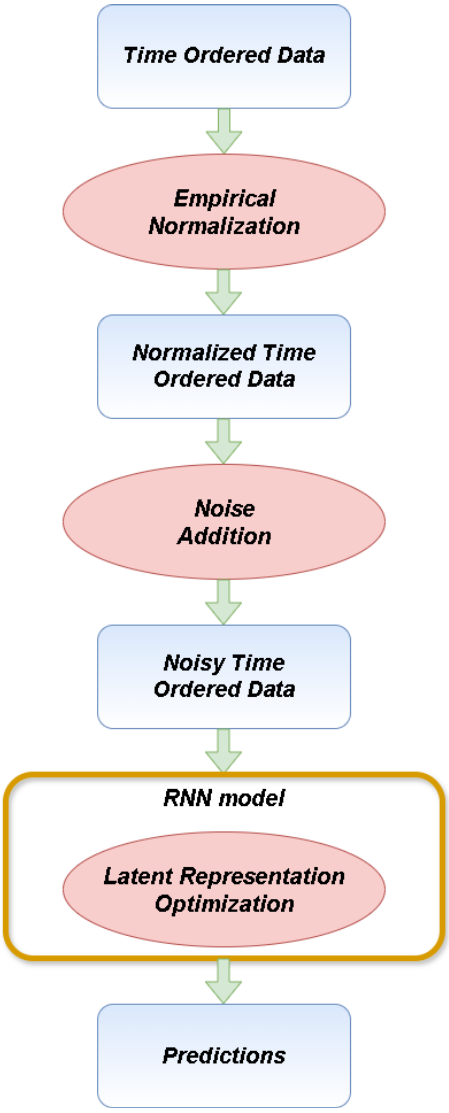

Figure 2.1: Framework of the proposed methodology for smart denoising optimization.

The first component of our methodlogy is Empirical Normalization, it standardizes the input data to zero mean and unit variance, which is essential for reducing overfitting risks. The second component is Latent Representation Optimization which focuses on optimizing latent representation by employing an autoencoder to compress the input data. This step enhances the model's ability to discern and learn fundamental data patterns, thus minimizing the impact of noise and variability.

The third component of our methodology is Noise Addition and involves the strategic addition of noise during the training process. This technique encourages the model to ignore random fluctuations and focus on learning robust, generalizable features. The success of our methodology is evidenced by substantial improvements in performance metrics and reductions in loss on unseen datasets.

The proposed methodology is visually summarized in Figure 2.1, illustrating the framework for smart denoising optimization. Although this chapter is relatively brief for such a crucial aspect of the thesis, we have deliberately divided the presentation of the proposed methodology and the evaluation framework into two separate chapters. This approach enhances clarity and organization, providing the reader with a more focused and insightful understanding of each aspect.

2.1 Empirical Normalization

Normalization techniques are crucial for preparing data before being used in model training, helping to ensure that the model works with data that is formatted or distributed in a way that optimizes performance. Popular methods for normalization include min-max scaling, z-score normalization, and adjusting data to fit a Gaussian distribution through different transformations. This process is especially important when dealing with noisy data or when data from various sources must be standardized to a uniform scale.

Empirical normalization is a crucial step in ensuring the successful implementation of the proposed method, as demonstrated in the results presented in Section 4.3, in the Figure 4.4 and Figure 4.10. Without empirical normalization, the method fails to produce accurate results. The rationale behind empirical normalization lies in the assumption that the added noise and the existing data share the same underlying distribution [17].

Two commonly used approaches to achieve empirical normalization are to either sample three-dimensional noise from the empirical distribution [18], [19], or to transform the input data into a Gaussian distribution using the empirical copula [20], [21] and subsequently sample Gaussian noise during the optimization process. Empirical studies have shown that normalization, regardless of its application to weights [22], batches [23], or data [24], contributes to improved stability and facilitates gradient descent.

2.2 Latent Representation Optimization

In deep learning optimization, it is common to start with weights that are assigned randomly. Studies have shown that backpropagation is dependent on initial weights selection [25] and using weights sampled from a normal distribution usually gives the best initial results [25]. Using the wrong distribution for these initial weights can cause problems such as the exploding gradient problem, where large error gradients accumulate and interfere with the training process. The main aim of Latent Representation optimization is to improve the starting point, helping the process converge more effectively towards the best solution. This approach is similar to what is done in autoencoder models, which simplify and encode data into a simpler format, thereby reducing noise and improving the representation of features right from the start.

The concept of Latent Representation Optimization is aimed at identifying an improved starting point for the optimization process [26]. Random selection of initial weights may not always lead to the optimal solution. Similar to the approach used in variational autoencoders [3], [27], the aim is to leverage the advantages of representing initial features in smaller spaces and compelling the model to disregard added noise.

However, the autoencoder component of the model becomes redundant after a certain point during the optimization process, as the representation becomes fixed as much as possible. It is ideal to disable the autoencoder component once this point is reached. To achieve this, two parameters, W_{ae} and $decay^{epoch}$, and a new term, are introduced into the loss function. W_{ae} is a weighting factor that determines the initial emphasis given to the autoencoder component, while $decay < 1$ determines the rate at which the optimization process transitions to focusing solely on the target variable. After $W_{ae} \cdot decay^{epoch}$ becomes sufficiently small, usually less than 0.25 in our case, the loss function is solely focused on the target variable to conserve computational resources.

The formulation of Latent Representation Optimization can be defined as:

$$L = L_Y + W_{ae} \cdot decay^{epoch} \cdot L_{ae}$$

Here, L_Y represents the supervised loss function that varies depending on the specific problem, while L_{ae} represents the loss between the encoded output and the input data.

While the latent representation optimization is not meant to be a standalone part of the optimization procedure, it can be evaluated as such. The results of whether it can serve as standalone optimization techniques are presented on the Figure 4.2 and Figure 4.8.

2.3 Noise Addition

The addition of noise is a well-established technique in deep learning [27] and has been shown to yield positive effects. By adding noise to the input data, we introduce variations that the model needs to account for, thereby improving the robustness and generalization capability of the model. The effectiveness of noise addition in deep learning is further evidenced by its successful implementation in generative adversarial networks [28].

The effectiveness of noise addition in denoising optimization can be achieved even without optimizing the latent representation, as demonstrated in Section 4.3 in Figure 4.1 and Figure 4.7. The primary objective of noise addition is to induce the model to recognize the intrinsic randomness of the stochastic process and learn to ignore it to capture the underlying nature of the process. By reducing overfitting, the model can achieve enhanced predictive power.

The algorithm for noise addition is presented in Algorithm 2.1, which involves three parameters: α, β, and sd. The parameters α and β control the decay rate of the added noise during training, with lower values corresponding to a faster decay in the amplitude of the noise and should be lower than 1 and higher than 0. In our analysis, we set β to 1 for simplicity, as its optimal value is highly dependent on the specific problem at hand. The parameter sd controls the amplitude of the noise, with lower values indicating a noise tensor closer to zero. The amount of noise added gradually decreases with each epoch to achieve the optimal balance between introducing variations and avoiding excessive noise that could hinder optimization. The following abbreviations are used in Algorithm 2.1:

- The input tensor is denoted by $X = [bs, ts, np]$, where bs, ts, and np represent the lot size, time steps, and the number of features used for predictions, respectively.
- The function mvn represents a multivariate normal distribution that generates samples from a two-dimensional correlated Gaussian distribution, where Σ represents the covariance matrix.
- The operator $matmul$ denotes the matrix multiplication operation.
- The abbreviation $chol$ refers to the Cholesky decomposition, a numerical method for factorizing a positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose. Similarly cov refers to function that computes the covariance matrix.

To improve the clarity of Algorithm 2.1, we provide a detailed explanation of its rationale. The objective is to incorporate a noise tensor with a three-dimensional structure that matches the probability distribution of the input dataset to facilitate learning during the optimization process. To accomplish this, the noise tensor is randomly generated and injected into the optimization loop after each epoch. This procedure ensures that the input

Algorithm 2.1: Noise addition.

Inputs: X, α, β, $epoch$, sd

Initialize matrix Y with dimensions time series length by time series length by number of parameters

$Y \leftarrow [bs, ts, np]$

Loop over each parameter to compute the Cholesky decomposition of the covariance matrix for each time slice

for $t \in \{1, \ldots, np\}$ **do**
  $\Sigma_t \leftarrow cov(X[,,t])$
  $Y[,,t] \leftarrow chol(\Sigma_t)$
end

Initialize matrix Z with dimensions batch size by time series length by number of parameters

$Z \leftarrow [bs, ts, np]$

Loop over each time series point to generate multivariate normal samples

for $i \in \{1, \ldots, ts\}$ **do**
  $\Sigma_i \leftarrow cov(X[,i,])$
  $Z[,i,] \leftarrow mvn(bs, \Sigma_i)$
end

Loop over each parameter to apply the Cholesky factor

for $j \in \{1, \ldots, np\}$ **do**
  $Z[,,j] \leftarrow matmul(Z[,,j], Y[,,j])$
end

Adjust Z matrix for decay effects over time and scaling

for $w \in \{1, \ldots, ts\}$ **do**
  $Z[,w,] \leftarrow Z[,w,] \cdot ((\beta^{ts-w} \cdot \alpha^{epoch}) \cdot sd)$
end

Add the noise matrix Z to the original matrix X to result in the final matrix R
  $R \leftarrow X + Z$
return R

stacks differ each time, enhancing the model's ability to generalize to diverse data scenarios. The algorithm's basic idea involves using copulas to establish the correct correlation structure within a multivariate normal distribution (mvn), followed by modifying the independent time dimension to match the desired correlation. To achieve this, two nested loops are employed. The first loop iterates through the indicators to capture their time correlation and stores the decomposition, while the second loop captures the correlation between the indicators for each time step. If the indicator's structure changes between the first and second loops, it must be included in the additional noise to maintain the correct correlation structure.

The Y variable epresents a series of Cholesky decompositions, one for each time step, where $Y[,t,]$ denotes the Cholesky decomposition of the covariance matrix for a given time step. While a sample includes noise with the correct correlation structure between indicators, the time dimension is still independent. The $Y[,t,]$ matrix corrects for this. Consider the example where the first and second indicators are related to long-term volatility. When generating correlated two-dimensional noise, both indicators will be similar. However, to accurately model the noise, the volatility indicators at time step T and $T-1$ must also follow the same structure. Thus, the noise structure must match the long-term volatility structure, which cannot change from one time step to the next.

After the second loop, a correlated noise tensor is created with the same dimensions as the input data, but the correlation between the time dimensions remains close to zero. This issue is addressed by multiplying the Cholesky decomposition of the time correlations with the simulated indicators. Finally, the noise is adjusted as a function of the current epoch of the training procedure. As the epoch number increases, the noise amplitude decreases, much like the *sd* parameter. Since the noise is generated with a mean of zero, it is added to the input data.

2.4 Preparing for Empirical Evaluation and Application

This section lays the groundwork for the empirical evaluation and practical application of the methodology introduced in this chapter, which focuses on smart denoising optimization. Chapter 3 introduces the evaluation methodology, detailing how each component of the proposed method will be assessed and statistically validated across a total of five different hypotheses. The separation of the evaluation methodology into its own chapter is essential, as it allows for a consistent application across both financial and textual data, supporting our scientific contributions. The hypotheses are detailed in Section 3.2.1, exploring whether noise addition contributes to superior performance or if the model performs better without it.

After setting up the evaluation framework, we detail the results in Chapter 4. Here, we evaluate and analyze each hypothesis, confirming or refuting them statistically. This analysis includes a full visualization of the parameter space to identify the best parameters for each dataset. In Chapter 5, we reinforce our findings with two real-world applications: one applies our methodology to a profitable trading strategy, and the other predicts the market demand for AI skills. In this way, we demonstrate that our methodology is useful and applicable in real-world contexts, particularly for intelligent denoising.

Chapter 3

Evaluation Methodology

In this chapter, we outline the evaluation methodology developed specifically to assess the effectiveness of the proposed methods under SC2 and SC3. A rigorous and structured approach is crucial to validate the capabilities and performance of our models in handling real-world data. We detail the various techniques and metrics employed to conduct a comprehensive analysis, ensuring that the findings are robust and reliable. This chapter serves as a pivotal link between the theoretical development presented in Chapter 2 and the practical applications demonstrated in subsequent sections. We have published the evaluation methodology and the results in the journal Information Sciences [16].

3.1 Structure of the Deep Learning Model

In recent years, recurrent neural networks (RNNs) have emerged as a promising approach for time-series prediction tasks, thanks to their ability to handle temporal dependencies in the data. Among the various types of RNNs, Long Short-Term Memory (LSTM) networks have been particularly successful due to their ability to capture long-term dependencies while avoiding the vanishing gradient problem that often plagues traditional RNNs [1]. In this context, a number of studies have demonstrated the superiority of LSTM models over traditional machine learning algorithms such as Random Forest and Generalized Linear Models (GLM) [14]. Specifically, it has been shown that LSTM models perform better on highly stochastic datasets, such as those related to cryptocurrencies and stocks.

Prior research has established the effectiveness of LSTM models in comparison to traditional machine learning approaches; however, to enhance the robustness of these findings, the present study was designed to increase confidence in the results by substantially increasing the number of iterations conducted. Each experiment was repeated 5000 times per dataset, necessitating a significant computational investment of thirty to sixty minutes of GPU time per stochastic stock dataset. This rigorous approach aimed to provide more compelling evidence for the superiority of LSTM models over conventional machine learning algorithms in time-series prediction tasks, building upon our prior work which demonstrated the advantage of the proposed LSTM model over GLMs, random forests, and simpler LSTM variants across two distinct datasets [14].

The denoising optimization technique proposed in this study is designed to be compatible with various recurrent deep learning structures. Given the documented superiority of Long Short-Term Memory (LSTM) units over traditional recurrent units or gate recurrent units [1], we decided to employ LSTM networks in our experiments. Our initial investigations revealed that the choice of model structure did not significantly impact the relative performance of the denoising optimization method. Specifically, while a better model structure may lead to increased performance, the denoising optimization approach consistently improved performance irrespective of the model architecture. However, the question of whether a faster model structure revision would yield better results remains an open one, which we intend to explore further in future research.

In our current study, we implemented a structure comprising three nested LSTM layers and a dense layer to construct the autoencoder. We also added three additional stacked dense layers to output the vector of the target variable(s). We used categorical cross-entropy as the loss function, appropriate for predicting the direction of trend changes in both datasets. The Adam optimizer [29] was chosen for model optimization. During our experiments, we varied several parameters to ensure comprehensive analysis, including the number of LSTM units (8, 16, 32) and dense units (8, 16, 32). We set the batch size at 2048 and programmed the models to halt training if there was no improvement on the validation dataset for 15 consecutive epochs.

3.2 Ablation Study and Statistical Analysis

Noise addition is a stochastic optimization technique that enhances machine learning model performance. It involves introducing random noise into the optimization process, which affects the outcomes, making them either favorable or unfavorable based on the specific evaluation criteria. To validate the reliability of results obtained through this method, we performed extensive testing, conducting 5000 evaluations per dataset across various parameters. The results section of our study is divided into three parts: an ablation study of the proposed method's different components, a statistical analysis demonstrating the method's superiority, and a visualization of the parameter space. While the ablation study was conducted separately for each of the five datasets, for statistical analysis and parameter space visualization, we combined datasets within the same domain, resulting in one dataset for the over 2000 financial domain datasets and one for the job posting skill demand dataset. The ablation study demonstrates outperformance based on specific metrics and loss functions, and this is further bolstered by statistical evidence supporting each hypothesis.

To examine the parameter space, we randomly selected various parameter combinations and evaluated them on the test set, ensuring unbiased parameter selection. We included all instances that met our criteria, even those that underperformed, to comprehensively analyze the hypotheses. For example, high noise amplitude might impede the model's convergence. We varied several parameters during the experiments, such as the number of LSTM units (8, 16, 32), dense units (8, 16, 32), the parameter α for noise addition (0, 0.5, 0.9, 0.99), with $\alpha = 0$ indicating no correlation and no noise reduction per epoch, the standard deviation sd in noise addition (0, 0.25, 0.5, 0.75, 1, 1.25, 1.5), where $sd = 0$ means no noise added, and the decay parameter in latent representation optimization (0, 0.25, 0.5, 0.75, 0.9), with $decay = 0$ signifying no optimization. The parameter combination of $\alpha = 0$, $sd = 0$, and $decay = 0$ was considered as the current state-of-the-art. The model structure was consistent, with nine combinations of unit numbers. We carried out a thousand evaluations per dataset for the non-stochastic state-of-the-art LSTM models. Our experiments were computationally intensive, requiring thirty to sixty minutes of GPU time per iteration on two NVIDIA GTX 2080Ti cards (11 Gb vRam each). Despite the computational demands, these extensive experiments provided us with robust and statistically significant evidence to support our findings.

3.2.1 Tested hypotheses:

The following hypotheses aim to evaluate the impact of noise addition and latent representation optimization on the performance of an LSTM model. We propose to investigate the following hypotheses:

Hypothesis 1: Optimization with noise addition outperforms standard LSTM

This hypothesis tests the effectiveness of noise addition as a denoising optimization method when compared to the current state-of-the-art LSTM model with the same structure. We define noise addition as an instance where the parameters $\alpha > 0$, $sd > 0$ and $decay = 0$. In contrast, the conventional LSTM model is trained with α, sd, and $decay$ set to 0.

Hypothesis 2: Latent representation optimization outperforms standard LSTM

This hypothesis evaluates the impact of latent representation optimization on the LSTM model's performance compared to the current state-of-the-art models. We add a penalty term to the loss function to focus on the latent representation, which is defined as an instance where the parameters $\alpha = 0$, $sd = 0$, and $decay > 0$. It is important to note that we implement a stochastic autoencoder to aid in the optimization process. We perform this analysis to understand the effects of latent optimization when noise addition is not included.

Hypothesis 3: Combining Latent representation optimization and noise addition outperforms the standard LSTM

In this hypothesis, we investigate the combination of noise addition and latent representation optimization and evaluate whether the performance of the LSTM model improves compared to the traditional optimization methods. The method selects a subset of test losses with all three active parameters: $\alpha > 0$, $sd > 0$, and $decay > 0$. The LSTM model is defined as in H_1 and H_2.

Hypothesis 4: Correlated noise addition outperforms uncorrelated non-decreasing noise addition

This hypothesis aims to determine whether adding correlated noise is necessary for the optimization process. We compare the performance of the LSTM model with noise addition to the model with uncorrelated, non-decaying noise addition (as in traditional variational autoencoders). We define uncorrelated, non-decreasing noise as an instance where the parameter $\alpha = 0$ and $sd > 0$.

Hypothesis 5: Latent representation optimization and noise addition outperform noise addition

This hypothesis evaluates whether the performance of the LSTM model benefits from the combination of latent representation optimization and noise addition compared to noise addition alone. We define noise addition alone as in H_1 and the combination of the two as in H_3.

3.2.2 Hypothesis evaluation

We split the 5000 test results into two subsets for each hypothesis and dataset, and we aim to determine whether there is a significant difference between the average test scores in these two datasets. To test for statistical differences, we use the Wilcoxon signed-rank test and visually compare the distributions. Table 3.1 specifies the number of scores tested for each hypothesis, which are the same for both datasets. Our aim is to draw conclusions from the analysis of each hypothesis and provide insights into the impact of noise addition and latent representation optimization on the LSTM model's performance. We aimed for 1000 samples per dataset, if the conclusion were clear sooner we have stopped evaluation earlier to save on compute.

	Sample 1	Sample 2
H_1 :	1000	1000
H_2 :	500	1000
H_3 :	1000	1000
H_4 :	1500	750
H_5 :	1000	1000

Table 3.1: Number of evaluations per hypothesis for both data sets. Sample 1 represents the first part of each hypothesis, while Sample 2 represents the second part, which corresponds to the LSTM models.

3.3 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a commonly used nonparametric method for data analysis [30]. Unlike parametric tests, which require specific assumptions about the data's distribution, the Wilcoxon signed-rank test does not assume a particular distribution shape, making it robust against outliers. This test is applicable for a single sample where the null hypothesis states that differences between paired observations are distributed symmetrically around zero [31]. The test ranks the absolute differences between paired observations, and the test statistic is derived from the sum of ranks for either positive or negative values [32].

One key advantage of the Wilcoxon signed-rank test is its flexibility with data that does not follow a normal distribution, which is a limitation in many parametric tests. It is especially effective for datasets with outliers or for small sample sizes. This test is widely applied across fields like medicine, psychology, and engineering, where it is used to compare two groups or assess the effects of treatments or interventions.

For unpaired data, the Wilcoxon signed-rank test has a counterpart—the Mann-Whitney test. In this version, the null hypothesis is that the two distributions differ only by a location shift of zero, while the alternative hypothesis suggests a different location shift. This non-paired version is appropriate when the observations are independent or unpaired.

In this thesis, we use the non-paired version of the Wilcoxon signed-rank test to compare the distribution of test losses across different hypothesis assumptions. Our objective is to reject the null hypothesis by achieving a low P-value. By applying a nonparametric test, we avoid distributional assumptions and ensure that our results are resistant to outliers. Overall, the Wilcoxon signed-rank test proves to be a reliable method for data analysis, particularly when dealing with non-normal distributions or small sample sizes [32].

3.4 Visualization of Parameter Space

To gain a deeper understanding of the performance of the denoising optimization method, we present a visualization of the impact of its parameters on performance. Since the addition of noise has been shown to be beneficial in both datasets, we focus on how different values of the parameters sd and α affect test performance. Specifically, we plot all 5000 instances with a heatmap, where each square on the heatmap represents the average test loss of all evaluations with specific values of α and sd. The parameter combination of $sd = 0$ and $\alpha = 0$ represents the state-of-the-art algorithm, as no noise has been added in this case. Instances with $\alpha = 0$ and $sd > 0$ correspond to experiments with uncorrelated, non-decreasing noise. The test losses are ranked, with lower rank indicating lower test loss. The heatmap's darker color indicates a lower rank, while the lower-left square represents the state-of-the-art and should be colored with lighter colors.

To facilitate the interpretation of the visualizations we provide a guide to the parameter space in Figure 3.1. The red color denotes the squares representing the average test loss of state-of-the-art models with no noise applied (i.e., sd set to 0). The gray color represents experiments where noise correlation is turned off by setting α to 0. The green area is the result of adding noise at different levels. Ideally, the green area should be the darkest in both figures.

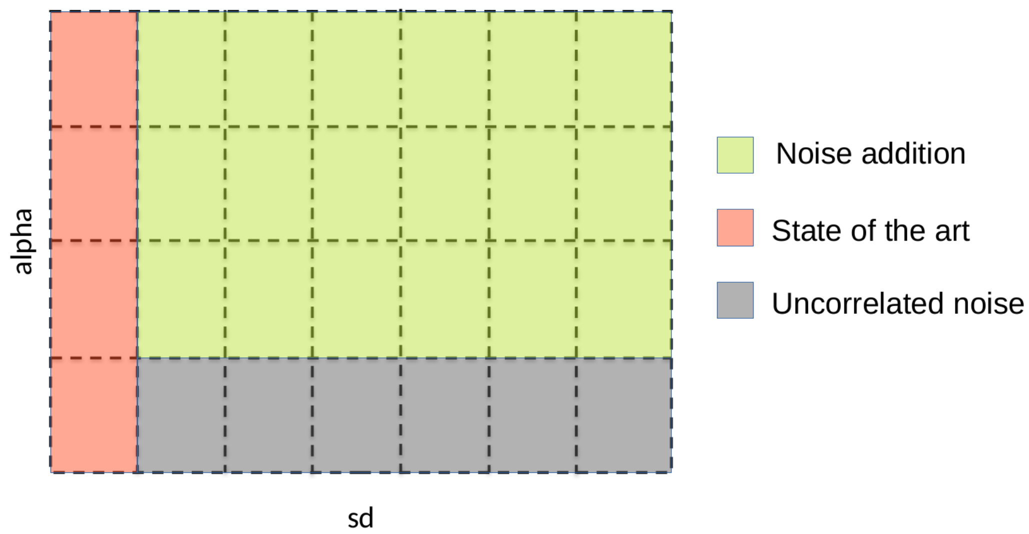

Figure 3.1: The guide for parameter space visualization of noise addition experiment.

Chapter 4

Empirical Analysis

In this chapter, we conduct a thorough empirical analysis to evaluate the effectiveness of our denoising optimization technique. The analysis spans multiple datasets from diverse domains, enabling us to test the robustness and adaptability of our methodology. We begin by outlining the data used in the study, including its source, structure, and relevance to our research objectives. Following this, we present an ablation study that systematically examines different components of our method, highlighting their individual and combined contributions to performance improvements. Finally, we undertake a detailed statistical analysis of the results to assess the statistical significance of our findings. This section includes a visualization of the parameter space, providing insights into the optimal settings for our denoising technique for both financial data and text data.

4.1 Data

To evaluate the effectiveness of our denoising optimization technique, we undertook an extensive analysis using a large and diverse collection of datasets. This data set included daily stock prices of 2,287 companies from diverse sectors like technology, banking, and real estate, spanning from 2002 to 2023. Additionally, we examined over 200 million job advertisements, categorized into 16 separate collections based on the country of the job posting.

Stock prices are theorized to adhere to a stochastic process influenced by Brownian Motion, as outlined in [33], while job postings reflect real-world skill demand. These postings are subject to fluctuations in supply and demand; however, labor market responses tend to be slower compared to the often rapid shifts seen in financial markets. Due to the sheer volume of over two thousand datasets, we categorized them into five distinct groups of unrelated datasets for practicality. Three of these groups belong to the financial sector, each representing different aspects of the global economy, and the remaining two groups comprise job postings, classified according to the language used in the listings. This categorization ensured a balanced representation of both English and non-English postings in each group, with around 100 million postings per language category. Our goal was to construct five independent datasets from varied fields, which would enable a robust demonstration of the effectiveness of our proposed methodology.

4.1.1 Financial Datasets

The financial domain datasets were sourced from Yahoo Finance's website [34] and analyzed using technical analysis techniques [35]. We segmented our data into two parts: data from 2002 to 2019 served as our training and validation set, while data post-2019 formed our test set. For validation method we used 5-fold cross validation. The dataset comprised 2,287 different stocks. Our primary focus was on predicting the direction of the three-day moving average of the closing price, both one and three days ahead. To improve the predictability of our model, we smoothed the target variable. This approach is necessary as the raw change in closing price over short intervals often defies prediction under the assumptions of Black-Scholes theory [33]. The data for each stock included Open, High, Low, Close, and Volume information. For each day at time T, we calculated 18 features using common technical indicators, including Bollinger Bands, various moving averages and oscillators. We applied standard parameters for these indicators. Our data representation involved six days of historical indicators, forming a two-dimensional matrix for each instance. Each stock's dataset is structured as [number of instances, time steps, number of technical indicators]. For clarity, we illustrate the format of a single instance in Table 4.1, where I indicates the value of the indicator and T represents a single day. On average, the price of a single stock is depicted through approximately 5000 matrices (accounting for non-trading days).

	I_1	$\cdots$	I_{18}
T	$x_{1,1}$	$\cdots$	$x_{1,18}$
$\vdots$	$\vdots$		$\vdots$
T-5	$x_{6,1}$	$\cdots$	$x_{6,18}$

Table 4.1: Data representation for a single day for a single process, where I represent indicator and T represent time.

We divided our analysis into three somewhat independent ensembles, each detailed in the following sections. To justify this split, we analyzed the correlation structure of the equally weighted closing prices of all stocks in each ensemble, as shown in Table 4.2. This correlation analysis confirms that the average stock from each dataset is largely independent from those in other datasets, suggesting that conducting separate experiments on each dataset can provide meaningful insights into the generalizability of our method. The datasets were also utilized in an ablation study to further evaluate our approach.

- Technological Dataset: This dataset comprises stochastic processes tracking the stock prices of companies like Microsoft, Apple, Amazon, and Tesla. These companies are typically correlated and benefit from a low-interest rate environment, with a focus on growth and innovation. This dataset consist of 1365 different process, totaling of roughly 6.8 million training instances.

- Industrial Dataset: This dataset includes stochastic processes monitoring the development of real estate (e.g., REITs), insurance and banking sectors, interest rate derivatives, and companies producing tools or weapons. This dataset consist of 534 different process, totaling of roughly 2.7 million training instances.

- Commodities Dataset: This dataset tracks stochastic processes related to the prices of various commodities, such as gold, steel, silver, copper, and companies dependent on these commodities. This dataset consist of 288 different process, totaling of roughly 1.5 million training instances.

- Equity Dataset: The equity dataset amalgamates the original 2287 stocks into a single, comprehensive dataset, which is later employed in statistical analysis and parameter visualization. This extensive dataset comprises over 10 million instances, providing a rich resource for in-depth analysis and exploration of stock market trends and behaviors.

	Technology	Industry	Commodities
Technology	1.00	0.17	0.12
Industry	0.17	1.00	0.31
Commodities	0.12	0.31	1.00

Table 4.2: Correlation values between datasets from the financial domain.

4.1.2 Text datasets

The skill demand dataset for our study was compiled from job postings across sixteen countries, collected from Adzuna between January 2019 and March 2022 [36]. To identify and extract skills or concepts from these postings, we utilized a tool known as Wikifier [37], which is capable of processing text in multiple languages. This tool generated time series data for each concept, illustrating its frequency of occurrence in job postings over time. The dataset's target variable is the future development of each concept in binary form, with instances after the first half of 2021 forming the test set. The total number of concepts analyzed is 5000. Similar to the stock dataset, we depicted the progression of each concept using technical indicators, formatted in the same three-dimensional structure as shown in Table 4.1. Despite the relatively short time span, this dataset encompasses millions of varied job postings, which have been aggregated into time series data for in-depth analysis. We further categorized the concepts based on geographical region, generating distinct time series for each country per concept. On average, each concept contributes approximately 200 training instances, implying that one can expect around 1 million training instances per region, though not every concept is necessarily present in each region. This rich dataset allows for a comprehensive examination of skill demand trends across different countries and time periods.

- Job Postings from English-Speaking Countries: This dataset predominantly features job postings from the United States, supplemented by those from Great Britain, collectively surpassing 100 million postings. These postings are analyzed and converted into time series data, totaling to roughly 1 million training instances.
- Job postings from Non-English Speaking Countries: This dataset includes job postings from countries such as Germany, India, France, Singapore, and Brazil, among others, again totalling to roughly 1 million training instances.
- Skill Demand dataset. In an approach analogous to that used in the financial domain, we aggregated all the data from the previously mentioned job postings dataset into a single comprehensive dataset for the purposes of statistical analysis and visualization. This unified Skill Demand dataset combines the various time series of concepts extracted from job postings across sixteen countries, totalling 2 million training instances.

The correlation of equally weighted time series for the concepts extracted from the job postings is detailed in Table 4.3. While the concept of averaging is more commonly associated with equities, particularly in the context of how exchange index funds are constructed, we applied a similar methodology to the job postings datasets. This approach was used to demonstrate the low correlation between the two datasets derived from job postings. By employing this method, we aim to highlight the distinct nature of each dataset, underscoring their individual characteristics and behaviors. This analysis is crucial in illustrating the diversity and independence of the datasets, supporting the validity of our approach in analyzing them separately.

	English	Non-English
English	1.00	0.28
Non-English	0.28	1.00

Table 4.3: Correlation values between datasets derived from job postings.

While the analysis of job postings offers valuable insights, it is not the core focus of this thesis. Instead, we utilize job postings data to establish an independent stochastic process, distinct from those observed in the financial sector. Our intention was to expand our analysis beyond the over 2000 financial processes, incorporating different types of stochastic phenomena. It is worth noting that the Geometric Brownian Motion, a pattern often seen in the financial domain, may not be as applicable to job postings, which are influenced by real-world skill demands.

Our methodology, originally developed for and primarily suited to financial applications [14], leverages the inherent randomness in stochastic processes. This research aims to explore whether our method can be effectively extended to other domains, such as skill demand. Unlike stock prices, the dynamics of skill demand are not wholly deterministic or completely stochastic. Prior studies have indicated limited success in applying denoising optimization techniques to non-stochastic datasets [15]. Therefore, this paper seeks to evaluate the applicability and efficacy of our approach in a domain that straddles the line between deterministic and stochastic processes, offering a novel perspective on the adaptability of financial models to broader contexts.

4.2 Ablation Study

In this section, we present the ablation study conducted on five datasets to evaluate the efficacy of the proposed method. The baseline for comparison was established with no noise addition and no latent representation optimization. We explored three distinct usage scenarios of our method: (1) optimization using only noise addition; (2) solely latent representation optimization; and (3) a combination of both, representing the complete proposed Denoising optimization method. Initially, accuracy was our primary metric. As we observed a direct inverse relationship between accuracy and categorical cross-entropy for the sake of consistency and clarity, we focus on the loss function, specifically categorical cross-entropy.

Our study involved 5000 evaluations for each of the four scenarios including the baseline. We selected all parameter combinations that fit each specific case. The baseline settings correspond to parameters α and sd set to zero, varying the number of LSTM and dense units. By examining both the best and worst performers in each scenario, we aimed to approximate the true mean outcome of the experiment. While using finely tuned examples could be representative, our approach is validated by the law of large numbers, suggesting that even without a distinct advantage, the best performer (in terms of hyperparameter settings) will appear favorable over time.

	Base	Noise Add.	Latent	Denoising Opt.
Technology	0.532	0.530	0.566	**0.528**
Industry	0.537	0.535	0.573	**0.533**
Commodities	0.534	0.532	0.551	**0.530**

Table 4.4: Comparison of average test losses on three financial datasets.

	Base	Noise Add.	Latent	Denoising Opt.
English skill demand	0.639	**0.638**	0.650	0.641
Non-English skill demand	0.642	**0.635**	0.650	0.643

Table 4.5: Comparison of average test losses on two skill demand datasets.

The results of the ablation study are summarized in Table 4.4 and Table 4.5, showing the average test losses across all experiments. It is evident that denoising optimization is beneficial for datasets from the financial domain (Technology, Industry, Commodities). However, its effectiveness for stochastic processes in the job market skill demand domain is less clear. One possibility is that the skill demand process is too deterministic, and denoising optimization introduces noise.

It may be that the stochastic process most suited to our proposed method should originate from the Brownian Motion family, mirroring the nature of the sampled noise. While it is promising to observe the method functioning as intended across various datasets from the original domain, relying solely on average comparisons is insufficient to assert this conclusively. Additional empirical evidence is required to support such a claim.

4.3 Statistical Analysis of the Proposed Method on Financial data

In this section, we present the results of statistical analysis on financial dataset. We utilized a color scheme to aid in the reader's comprehension of the outcomes. The densities of the noise addition component of our method are denoted in blue, the standard LSTM model in red, and the green represents both the noise addition and latent optimization plots. The grey colour corresponds to densities that are considered less important because they should not be used independently, specifically uncorrelated noise addition and latent optimization representations. It should be noted that all models utilized the same structure, and the only variation between the two samples studied was the optimization procedure.

The results of our study indicate that incorporating noise addition and latent optimization enhanced the performance of the standard LSTM model. The blue density plots, which represent the noise addition component, demonstrate that the incorporation of noise addition aided in reducing overfitting. The standard LSTM model in red showes a higher level of overfitting in comparison to the noise addition and latent optimization plots in green. This suggests that the addition of noise can aid in enhancing the model's generalizability.

We approach our analysis by examining each hypothesis in detail, confirming the statistical significance of the findings from the ablation study. This involves both visual comparisons and numerical assessments. Instead of comparing all five datasets individually, which would be cumbersome and less informative, we have consolidated the datasets from each category into two large datasets for more comprehensive statistical analysis. For each hypothesis, we present a table displaying the test losses for the relevant experiment, the relative improvement observed, and the Wilcoxon test p-value. The latter is crucial for determining the statistical significance of the differences between the two models, with a smaller p-value indicating higher confidence in the difference being significant.

Under each hypothesis, we will discuss the results, providing a thorough interpretation of what these findings signify. This structured approach aims to present the statistical analysis in a clear and logical manner, enabling the reader to grasp the implications of our study's outcomes effectively.

4.3.1 Hypothesis evaluation

Hypothesis 1: Optimization with noise addition outperforms standard LSTM.

This experiment investigates the impact of optimization with noise addition on test performance compared to the current state-of-the-art LSTM model. The average test losses for both the noise addition and LSTM models are presented in Table 4.6. In addition, the Wilcoxon test p-value is provided to assess the statistical significance of the difference between the two models. A smaller p-value indicates a greater confidence in the difference being greater than zero. The results in Table 4.6 demonstrate that the addition of noise improves test performance for both equity and skill demand datasets. The improvement in test loss is more significant for the equity dataset than for the skill demand dataset. This could indicate either that denoising is more effective for certain stochastic processes or that the dataset for demand for skilled labor is not large enough to accurately estimate the underlying distribution.

	Noise Addition	LSTM	Improvement (%)	P-value
Equity	0.5334	0.5364	0.56	$2.2e^{-16}$

Noise Addition	LSTM	Improvement (%)	P-value

Table 4.6: Comparison of average test losses, relative improvement, and Wilcoxon test p-value of the noise addition model to the current state-of-the-art LSTM model.

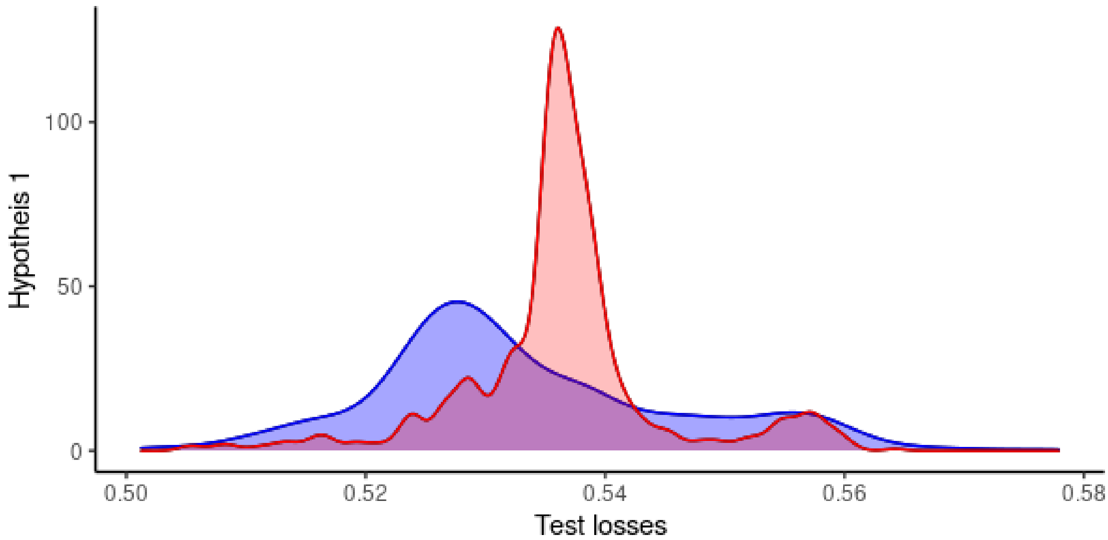

Figure 4.1: Density comparison of test losses between the noise addition (blue) and standard LSTM model (red).

A novel dimension of our analysis is introduced through the visualization of the density distributions. The baseline models, depicted in red, encompass all hyperparameter configurations where both noise addition and latent optimization were inactive. The resulting density displays a single dominant peak with minimal heavy tails, indicating a concentration of results around a specific region. In contrast, the blue densities, representing the noise addition model, exhibit heavier tails and a leftward skew, indicating smaller test losses. It is crucial to highlight that these densities are not selectively drawn from the optimal hyperparameter settings; they also include less successful configurations. This comprehensive dataset suggests that with meticulous fine-tuning on training and validation sets, one could potentially achieve superior performance on unseen data beyond what the relative improvement currently indicates.

Hypothesis 2: Latent representation optimization outperforms standard LSTM.

This hypothesis focuses on assessing the individual contribution of latent representation optimization. While the previous hypothesis highlighted the impact of noise addition, we now explore whether latent representation optimization alone can also lead to performance improvements. The results of this investigation are detailed in Table 4.7 and visually represented in Figure 4.2. The findings indicate that when applied in isolation, latent representation optimization adversely affects performance on test data.

	Latent Opt.	LSTM	Improvement (%)	P-value
Equity	0.5385	0.5364	-0.4	0.6351

Table 4.7: Comparison of of test losses of Latent Optimization and baseline model.

To gain a deeper understanding of the results, we plot the density comparison of test losses for both datasets in Figure 4.2. Grey densities in the plot represent the performance of models using only latent representation optimization. A divergence in behavior is observed in the equities dataset. Although the decrease in performance is not statistically significant and could be attributed to randomness, the density plot reveals potential benefits. The primary peak has shifted significantly to the left, indicating lower test losses, but a secondary peak has also emerged on the right tail of the distribution, with a heavier left tail observed as well. This suggests that the combination of latent representation optimization with noise addition might be effective in filtering out the underperforming right peak, thereby harnessing the latent representation optimization's potential.

The variance in outcomes likely stems from the nature of the underlying processes. It appears that datasets characterized by heavily stochastic processes, such as those following Geometric Brownian Motion, are more amenable to the proposed method. This insight suggests that the effectiveness of latent representation optimization, particularly when combined with noise addition, may be contingent on the specific characteristics of the dataset being analyzed.

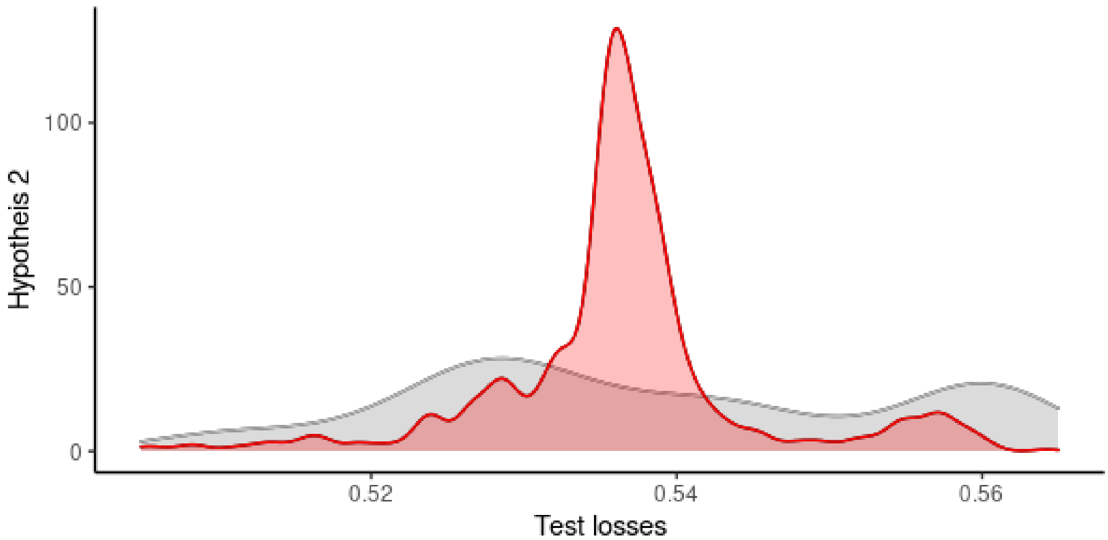

Figure 4.2: Density comparison of test losses of only the latent representation optimization (grey) and standard LSTM model (red).

Hypothesis 3: Combining Latent representation optimization and noise addition outperforms the standard LSTM.

This hypothesis is crucial to this research as it examines whether the combination of both components of our method enhances performance on test datasets. The findings are detailed in Table 4.8, with corresponding density plots shown in Figure 4.3. In these plots, the green densities represent our proposed method, while the red densities signify the benchmark. For the equities datasets, our method demonstrates remarkable improvement, supporting the hypothesis with a P-value below 0.01. The results in Table 4.8 indicate a significant enhancement in test set performance when denoising optimization and latent representation optimization are combined. This combination effectively mitigated noise and provided a more accurate approximation of the underlying process. Figure 4.3 visually compares the distribution of test losses between our method and current state-of-the-art models, highlighting the substantial improvements achieved.

Previous results confirmed that adding noise is beneficial, as indicated by our validation of the first hypothesis. When both noise addition and latent representation optimization are applied to the equities datasets, we observe a promising shift in performance. Specifically, the method helps reduce the right peak and skews results towards the left tail in the distribution of losses on unseen data, suggesting significant potential for improvement through hyperparameter tuning. This evidence suggests that the combination of noise addition and latent representation optimization may be particularly effective for the equities datasets. The results indicate a significant enhancement in model performance, underscoring the potential of our method to achieve lower test losses with appropriate fine-tuning.

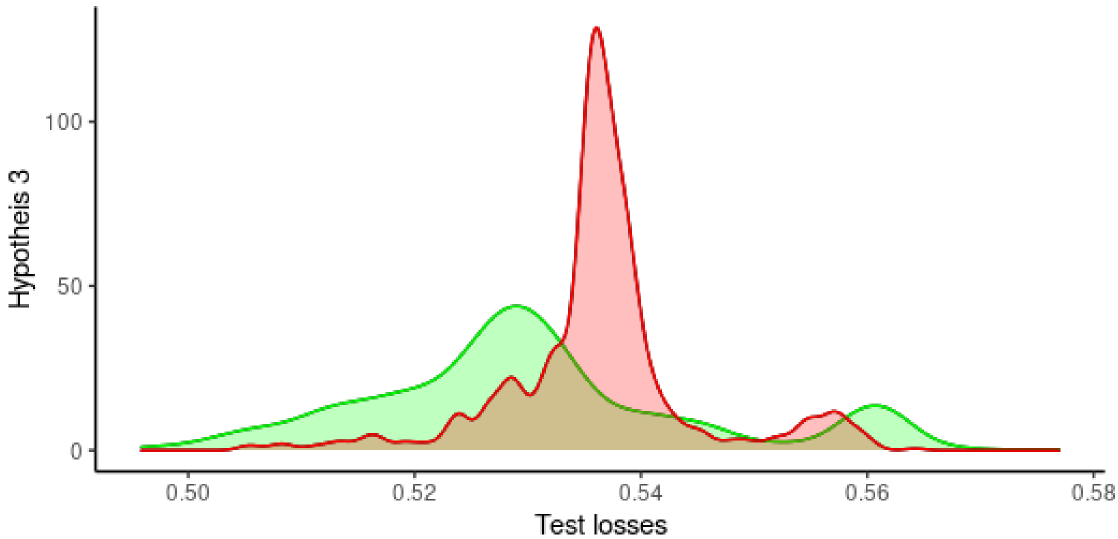

Figure 4.3: The density comparison of test losses between the denoising optimization (green) and the standard LSTM model (red).

	Denoising opt.	LSTM	Improvement (%)	P-value
Equity	0.5305	0.5364	1.10	$2.2e^{-16}$

Table 4.8: Comparison of average test losses between the denoising optimization and the current state-of-the-art LSTM, with relative improvement and P-value of the Wilcoxon test to determine statistical significance.

The present study has successfully completed a thorough comparison between the proposed methodology and the current state of the art. The empirical results suggest that the proposed method exhibits significant outperformance when applied to equities data.

Furthermore, the experimental findings indicate that the addition of noise can enhance the performance of the proposed method for the equities dataset. To delve deeper into the issue of noise correlation and its necessity in the proposed method, subsequent tests will be conducted to scrutinize the impact of noise-only addition on the overall performance of the model. Additionally, another set of experiments will assess whether there exists a significant difference between the overall approach and the noise-only addition.

Hypothesis 4: Correlated noise addition outperforms uncorrelated non-decreasing noise addition.

In this hypothesis, our objective is to demonstrate that incorporating noise correlation, despite its computational demands, is essential for the efficacy of our method. The results of this investigation are summarized in Table 4.9, with corresponding density plots displayed in Figure 4.4.

Applying a simpler and faster optimization process would have been highly beneficial from a practical standpoint. However, both numerical and visual analyses clearly indicate the necessity of noise correlation for improved performance across datasets, thus confirming the hypothesis.

	Noise Add.	Uncorrelated noise	Improvement (%)	P-value
Equity	0.5316	0.5568	4.73	$2.2e^{-16}$

Table 4.9: Comparison of average test losses of the noise addition to uncorrelated non-decreasing noise, relative improvement, and P-value of Wilcoxon test whether the difference is statistically significant.

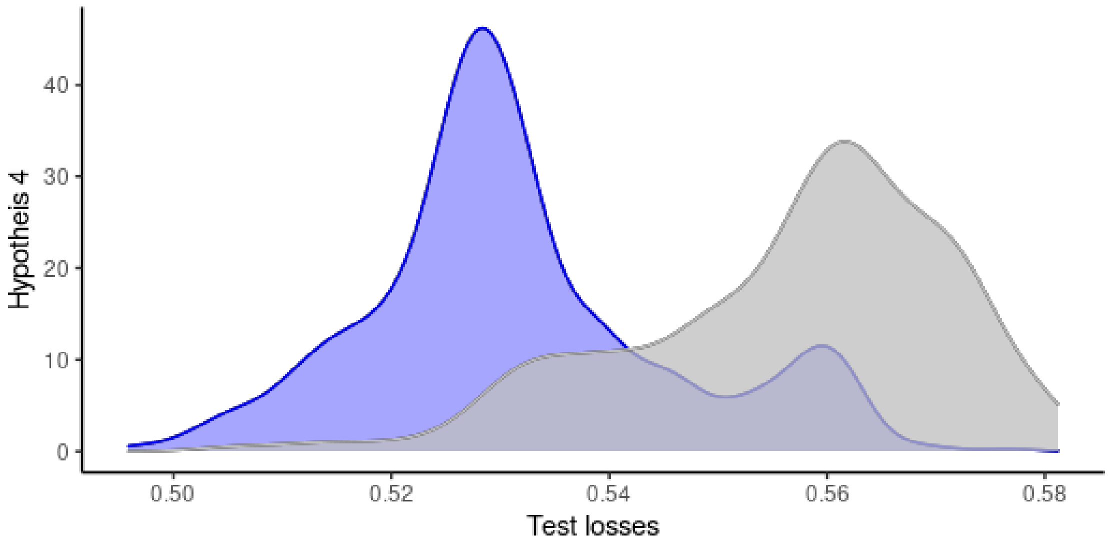

Figure 4.4: Density comparison of test losses of the noise addition (blue) and uncorrelated non-decreasing noise (grey).

The difference in densities with and without noise correlation is striking, with the former showing a substantial enhancement. This outcome is not entirely unexpected.

The incorporation of a correlation structure provides deeper insights into the underlying process, enabling the method to discern between noise and critical signals in the data. This distinction is crucial for accurately interpreting and processing the information.

Hypothesis 5: Latent representation optimization and noise addition outperform noise addition.

The experimental results presented in Table 4.10 and Figure 4.5 demonstrate the efficacy of incorporating latent optimization in the proposed model. The observed density plots for the equity dataset reveal that the inclusion of latent optimization (represented by the green plot) leads to a thicker left tail, indicating a significant improvement in the model's performance with respect to lower test losses. Although the peaks have shifted to the right, the observed pattern emphasizes the sensitivity of the model's performance to parameter choices. The results clearly highlight that the selection of appropriate noise and reduction values is crucial in achieving the desired test performance.

	Denoising opt.	Noise addition	Improvement (%)	P-value
Equity	0.5305	0.533	0.47	0.0005

Table 4.10: Comparison of average test losses of denoising optimization to noise addition, relative improvement, and P-value of Wilcoxon test whether the difference is statistically significant.

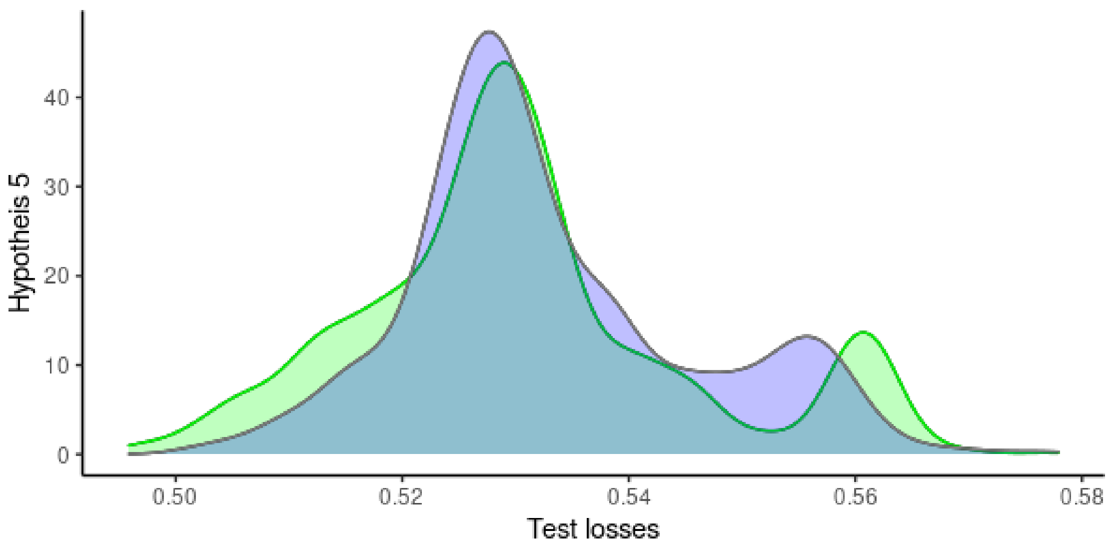

Figure 4.5: Density comparison of test losses of the denoising optimization (green) and noise addition (blue).

Table 4.10 provides a comprehensive comparison of the average test losses between denoising optimization and noise addition, along with the relative improvement and the P-value of the Wilcoxon test to determine the statistical significance of the difference. The experimental results reveal a reduction in the average test loss for the equity dataset when the denoising optimization is used. The observed trends demonstrate the effectiveness of the proposed denoising optimization approach in enhancing the model's overall performance. The significant difference in test performance between the two methods emphasizes the importance of selecting an appropriate optimization technique.

Figure 4.5 provides a visual comparison of the density plots of test losses for the denoising optimization and noise addition methods. The green and blue density plots represent the denoising optimization and noise addition methods, respectively. The observed multiple peaks in most density plots indicate that certain parameter choices lead to better model performance. The experimental results suggest that the optimal parameter selection requires a careful balance between the amount and amplitude of added noise to ensure the model effectively learns from input data while still being robust to noise.

4.3.2 Visualization of parameter space

Figure 4.6 features a heatmap that illustrates the impact of the parameters α and sd on the equity datasets when using the denoising optimization method. This heatmap compiles data from 5000 instances, with each square representing the average test loss for particular values of α and sd. The color intensity of each square correlates with the test loss rank, where darker colors signify a lower (and thus more favorable) test loss rank. Notably, the combination of $sd = 0$ and $\alpha = 0$—which is the current state-of-the-art algorithm with no noise addition—occupies the lowest rank in the lower left corner.

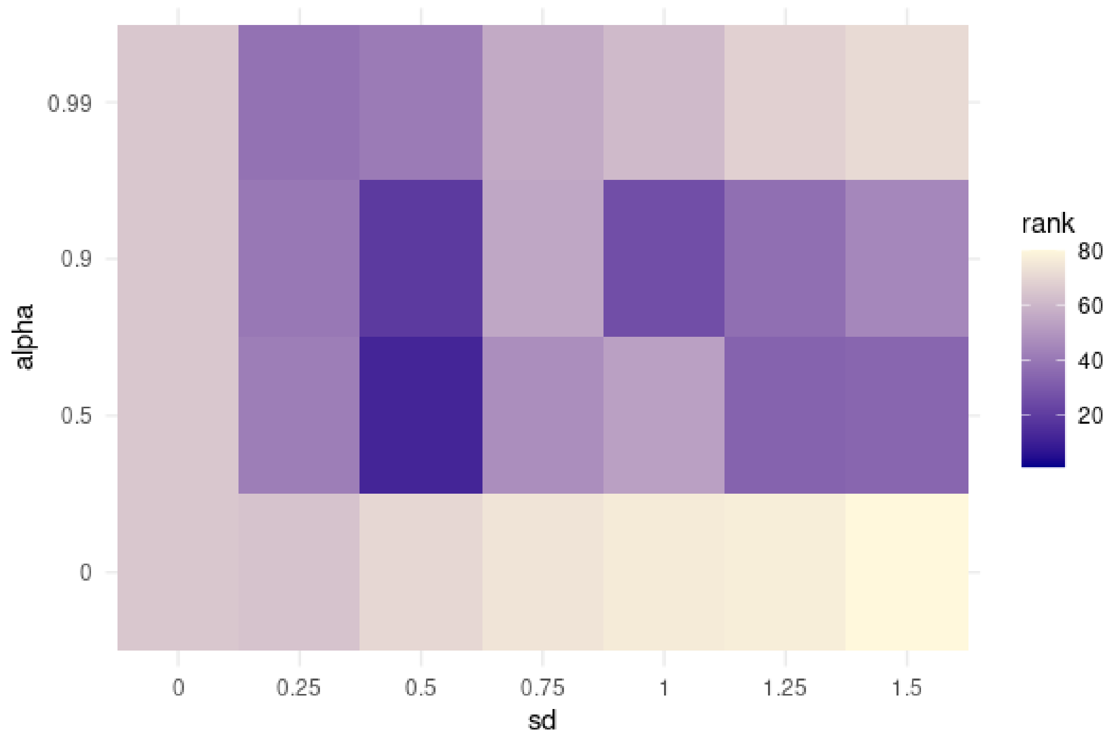

Figure 4.6: Visualization of parameters α and sd from noise addition on equity dataset.

The results demonstrate the beneficial effect of noise addition in the denoising optimization process. A particularly effective parameter combination, identified by the darkest area on the heatmap, is at $\alpha = 0.5$ and $sd = 0.5$. This represents an optimal balance for the method's performance. While the combination of $\alpha = 0.9$ and $sd = 0.75$ shows less consistent results, potentially due to the stochastic nature of optimization, there is still an overall trend of continuity in the heatmap, reinforcing the efficacy of the method.

Furthermore, the hypothesis tests also show that noise correlation is necessary for the denoising optimization method. These findings indicate that the denoising optimization method can outperform the current state-of-the-art algorithm by adding noise to the dataset. Therefore, these results provide strong evidence that adding noise to the dataset is an effective strategy for improving the performance of the denoising optimization method.

4.4 Statistical Analysis of the Proposed Method on Text Data

In this section, we shift our focus to evaluating the results of our methodology on job data. This dataset presents a unique set of challenges and opportunities, offering valuable insights into the performance of our approach in a distinct domain. Similar to our analysis on financial data, we employed a color scheme to facilitate the interpretation of outcomes. Blue denotes the densities of the noise addition component of our method, red represents the standard LSTM model, and green illustrates both the noise addition and latent optimization plots. Additionally, grey indicates densities deemed less significant due to their dependence on correlated noise addition and latent optimization representations. It is important to emphasize that all models adhered to the same structure, with the only variation being the optimization procedure.

Our study's results suggest that incorporating noise addition and latent optimization enhances the standard LSTM model's performance. The blue density plots, representing the noise addition component, indicate that incorporating noise aids in mitigating overfitting. In contrast, the standard LSTM model in red exhibits a higher level of overfitting compared to the noise addition and latent optimization plots in green. This observation suggests that noise addition contributes to improving the model's generalizability. Subsequently, we will delve into discussing the results under each hypothesis, offering a thorough interpretation of their implications. This structured presentation aims to elucidate the statistical analysis in a coherent and comprehensible manner, enabling readers to discern the significance of our study's outcomes effectively.

4.4.1 Hypothesis evaluation

Hypothesis 1: Optimization with noise addition outperforms standard LSTM.

This experiment delves into evaluating the impact of optimization using noise addition on test performance, juxtaposing it against the current state-of-the-art LSTM model, particularly in the context of job demand dataset. Table 4.11 presents the average test losses for both the noise addition model and the LSTM model, complemented by a density visualization in Figure 4.7.

	Noise Addition	LSTM	Improvement (%)	P-value
Skill demand	0.6382	0.6408	0.41	$1.42e^{-15}$

Table 4.11: Comparison of average test losses, relative improvement, and Wilcoxon test p-value of the noise addition model to the current state-of-the-art LSTM model.

Analyzing the components of our proposed method, noise addition emerges as a pivotal factor. Notably, noise addition demonstrates statistical significance on the job demand dataset, consistent with findings from the earlier ablation study. This affirmation of the hypothesis for the job demand dataset, with a P-value below 0.01, underscores the significance of noise addition in enhancing model performance within this domain.

The visualization of densities in the job demand dataset sheds new light on our analysis. Unlike the Equity dataset, which exhibits a single dominant peak with minimal heavy tails, the job demand dataset displays a distinct bimodal distribution. The blue densities, representing the noise addition model, exhibit heavier tails skewed towards the left, indicative of smaller test losses. It's crucial to note that these densities encompass various hyperparameter settings, including both successful and less successful ones. This diversity

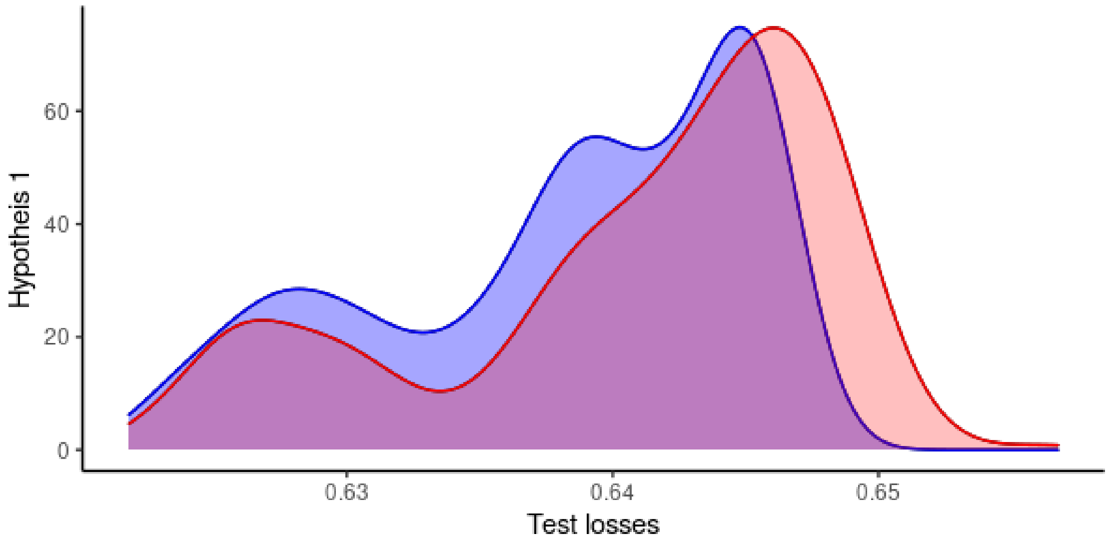

Figure 4.7: Density comparison of test losses between the noise addition (blue) and standard LSTM model (red) on job posting data. The figure visually illustrates the performance improvement of the noise addition model on unseen data.

suggests that further refinement through appropriate fine-tuning on training and validation sets could potentially yield even greater performance enhancements on unseen data than initially indicated by the relative improvement.

Hypothesis 2: Latent representation optimization outperforms standard LSTM.

This hypothesis explores the isolated impact of latent representation optimization specifically within the job demand dataset. As detailed in Table 4.12 and visually represented in Figure 4.8, our analysis reveals a significant decrease in performance, with a P-value below 0.01. Grey densities in the plots distinctly illustrate this disadvantage, showcasing a shift towards higher test losses and a reduction in the left heavier tail. This observation highlights the adverse effect of latent representation optimization when applied independently within the job demand dataset. The variance in outcomes suggests that the effectiveness of this optimization strategy may be contingent upon the specific characteristics of the dataset being analyzed, particularly within the context of job demand data.

	Latent Opt.	LSTM	Improvement (%)	P-value
Skill demand	0.6503	0.6408	-1.46	$2.2e^{-16}$

Table 4.12: Comparison of of test losses for job posting data. The figure highlights the state-of-the-art LSTM models in red and the models with latent representation optimization in grey.

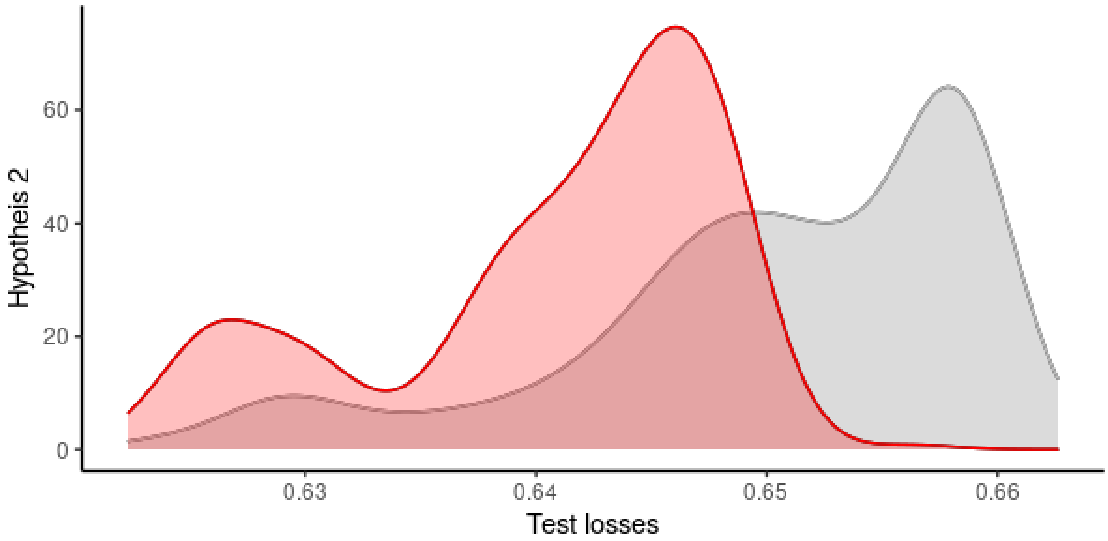

Figure 4.8: Density comparison of test losses of only the latent representation optimization (grey) and standard LSTM model (red).

Hypothesis 3: Combining Latent representation optimization and noise addition outperforms the standard LSTM.

This hypothesis critically examines whether combining both components of our method enhances performance on the skill demand dataset. The detailed findings are presented in Table 4.13, accompanied by corresponding density plots in Figure 4.9. For the skill demand dataset, our analysis reveals that the second component of our method, namely the latent representation, acts as a limiting factor. The proposed method, in its current form, fails to improve results with a P-value below 0.01. This contrasts with the remarkable improvement observed in equities datasets, affirming the hypothesis.

	Denoising optimization	LSTM	Improvement (%)	P-value
Skill demand	0.6424	0.6408	-0.25	$3.6e^{-6}$

Table 4.13: Comparison of average test losses between the denoising optimization and the current state-of-the-art LSTM, with relative improvement and P-value of the Wilcoxon test to determine statistical significance.

A deeper investigation into the suboptimal performance on the skill demand dataset is warranted. Figure 4.12 in the subsequent chapter provides a granular breakdown of the parameter space, revealing that only low α values and minimal noise levels yield improved predictions on the test dataset. Notably, this region of improvement is significantly narrower compared to the equity dataset, indicating a lack of overall improvement when results are averaged. This discrepancy is likely attributed to the skill demand dataset's lower level of inherent randomness, posing challenges to the method's efficacy designed for datasets characterized by high uncertainty.

Figure 4.9: The density comparison of test losses between the denoising optimization (in green) and the standard LSTM model (in red).

Hypothesis 4: Correlated noise addition outperforms uncorrelated non-decreasing noise addition.

In this hypothesis, our aim is to demonstrate the essential role of noise correlation in enhancing the efficacy of our method within the job demand dataset. The summarized results of this investigation can be found in Table 4.14, with corresponding density plots displayed in Figure 4.10. While a simpler and faster optimization process would have been advantageous from a practical standpoint, both numerical and visual analyses unequivocally emphasize the necessity of noise correlation for improved performance across datasets, particularly within the job demand dataset, thus validating the hypothesis.

	Noise addition	Uncorrelated noise	Improvement (%)	P-value
Skill demand	0.6382	0.6599	3.40	$2.2e^{-16}$

Table 4.14: Comparison of average test losses of the noise addition to uncorrelated non-decreasing noise, relative improvement, and P-value of Wilcoxon test whether the difference is statistically significant.

The stark difference in densities between scenarios with and without noise correlation is notable, with the former showcasing significant enhancement in both datasets. This observation underscores the importance of incorporating a correlation structure, which provides deeper insights into the underlying process. Specifically, it enables the method to discern between noise and critical signals in the data, a crucial distinction for accurate interpretation and processing of information.

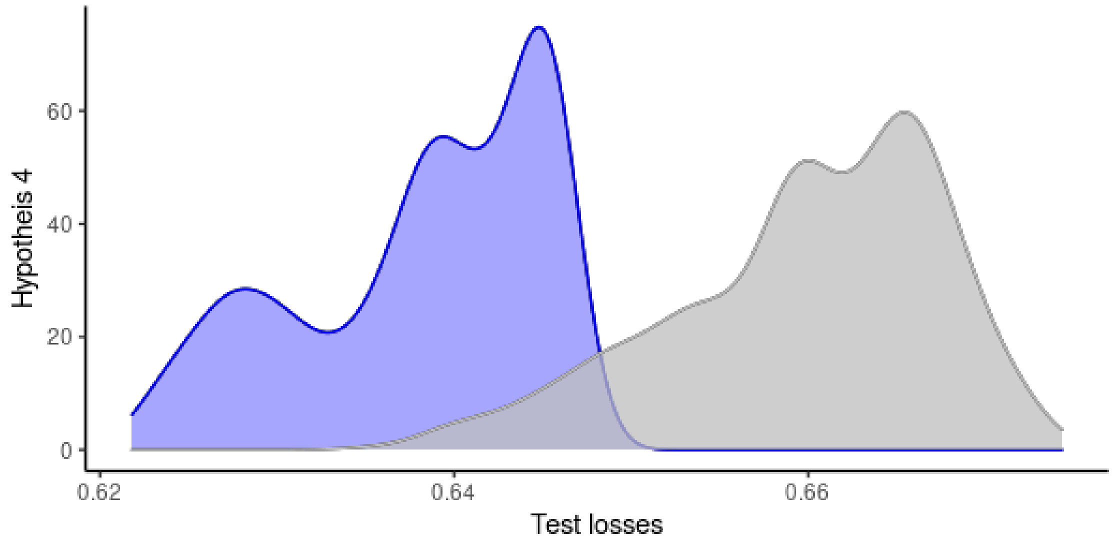

Figure 4.10: Density comparison of test losses of the noise addition (blue) and uncorrelated non-decreasing noise (grey).

Hypothesis 5: Latent representation optimization and noise addition out-perform noise addition.

This hypothesis aims to discern the statistical differences between adding noise alone and implementing the full method, particularly within the job demand dataset. While it may be considered less critical among the tested hypotheses, its focus on equity datasets and visual analysis of density plots provides valuable insights. Previous findings have established no improvement in the skill demand dataset.

	Denoising optimization	Noise addition	Improvement (%)	P-value
Skill demand	0.6424	0.6382	-0.66	$2.2e^{-16}$

Table 4.15: Comparison of average test losses of denoising optimization to noise addition, relative improvement, and P-value of Wilcoxon test whether the difference is statistically significant.

Results in Table 4.15 and Figure 4.11 reveal straightforward outcomes for the job demand dataset. The latent representation component predominantly influences the outcomes, shifting the peak rightward and flattening the left tail, resulting in a double disadvantage. However, within the job demand dataset, a notable thickening of the left tail suggests enhanced performance on unseen data. Interestingly, there is potential for further reduction of the right peak. This implies that with appropriate fine-tuning, it might be possible to optimize the method's efficacy further.

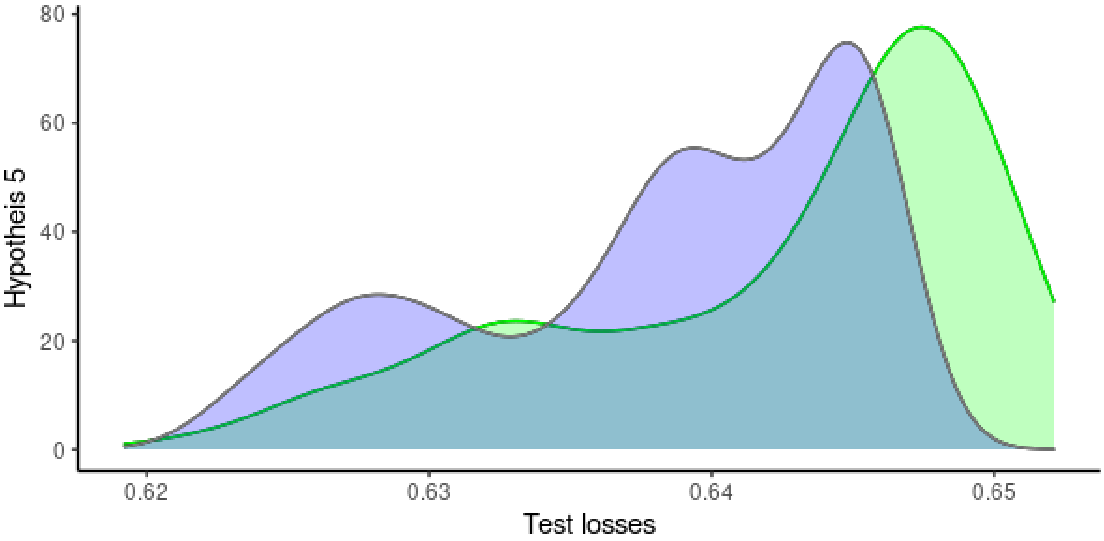

Figure 4.11: Density comparison of test losses of the denoising optimization (green) and noise addition (blue).

4.4.2 Visualization of the parameter space

Figure 4.12 serves as a focal point for assessing the performance of the denoising optimization method within the context of the job demand dataset. The visualization offers insights into how variations in parameters α and sd influence the effectiveness of the method. Examining the heatmap, it becomes apparent that lower values of α and sd tend to yield more favorable outcomes. This observation suggests that the job demand dataset, characterized by its deterministic nature, responds better to subtle perturbations introduced by lower levels of noise. Conversely, regions with excessively high noise levels exhibit diminished effectiveness, indicating a critical threshold beyond which noise ceases to enhance the method's performance.

Comparing these results to those of the state-of-the-art algorithm without noise, we observe comparable brightness levels in regions where the denoising optimization method is applied. This parity underscores the method's efficacy in mitigating the effects of noise and optimizing performance within the job demand dataset. These findings underscore the importance of parameter fine-tuning and noise incorporation in dataset optimization. By carefully adjusting parameters and introducing controlled noise, researchers can harness the full potential of the denoising optimization method, thereby enhancing its performance across various datasets, including those characterized by deterministic patterns like the job demand dataset.

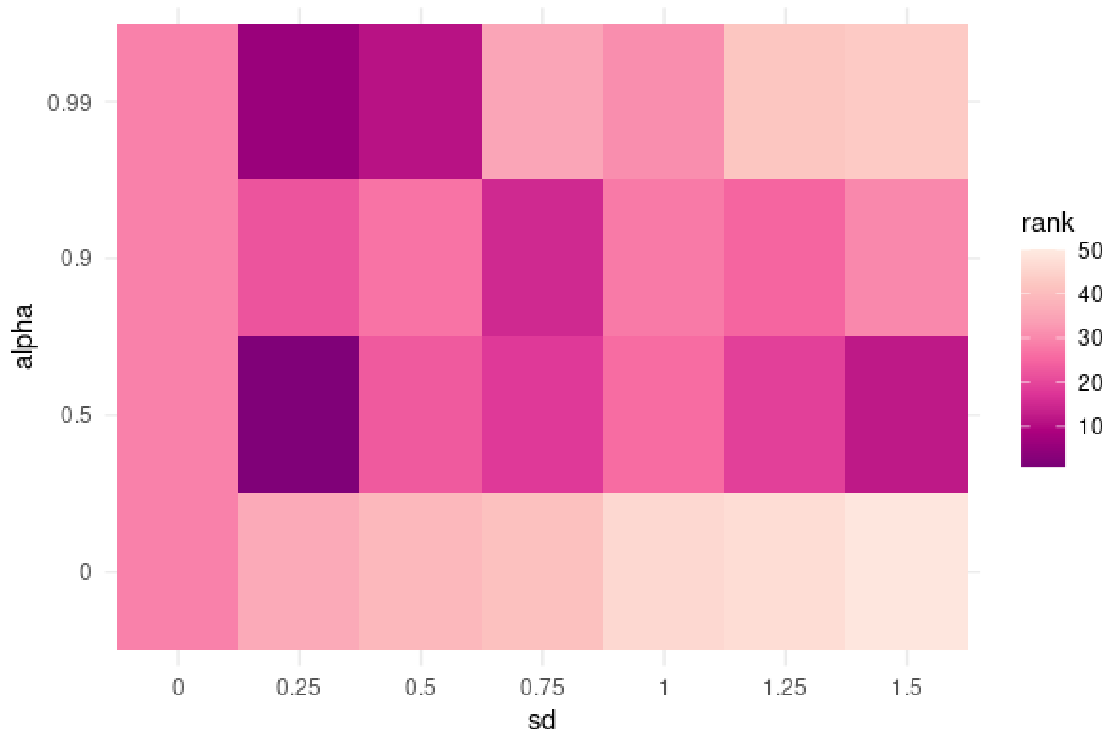

Figure 4.12: Visualization of parameters α and sd from noise addition on skill demand dataset.

Chapter 5

Application of the Proposed Methodology on Real-World Use Cases

In this section of the dissertation, we transition from the theoretical foundations to the practical application of our proposed methodology.

The primary goal at this juncture is to assess the utility of our approach within actual trading environments, thereby validating its effectiveness in the dynamic realm of financial markets. To establish a comprehensive baseline, we have selected a trading strategy that is distinguished by its positive expected value and minimal volatility. This strategy will serve as a benchmark for evaluating the advantages of our methodology. We will undertake a detailed examination of this baseline strategy, elucidating its complex mechanisms and underlying principles.

In the subsequent example, we demonstrate the application of our methodology to a real-world case derived from text data, where we model AI demand across different countries. This example illustrates how our approach can enhance predictive accuracy in practical scenarios: Job demand prediction [38]. A reference to this application was detailed in a publication by the IEEE Access journal (Impact Factor: 3.9, CiteScore: 9, 2024).

5.1 Use Case on Profitable Trading Strategy

Important practical test for our methodology lies in its integration with a real word scenario, such as profitable trading strategy. We will strategically deploy our predictive model to identify opportune moments for abstaining from entering into the baseline strategy. By tactically sidestepping trades during specific market conditions, we posit that the strategy's already favorable risk-reward profile can be further augmented. To quantify the potential enhancements, we will conduct a meticulous analysis of key performance metrics.

With the aim of closely mirroring real-world trading scenarios, we will disclose a trading strategy that we ourselves have actively traded, involving the utilization of 0 DTE (Days to Expiration) options on the SPX index within the US equity markets. Before delving into the specifics of this strategy, it's essential to elucidate some fundamental concepts surrounding options trading. An options contract grants the holder the right, but not the obligation, to buy or sell an underlying asset at a predetermined price (known as the strike price) within a specified period of time. These contracts are traded on various financial instruments, including stocks, indices, and commodities, and they are widely utilized by investors and traders for various purposes, such as hedging against price fluctuations, generating income through premiums, or speculating on market movements.

In our trading strategy, we focus on 0 DTE options, which stands for "zero days to expiration." This term refers to options contracts that are set to expire at the end of the trading day, providing traders with a unique opportunity to capitalize on short-term price movements. 0 DTE options are characterized by their short expiration period, typically spanning from the time of purchase to the market close on the same day. As such, they are particularly suited for day trading strategies that capitalize on intraday volatility and price fluctuations. Trading 0 DTE options on the SPX index entails selecting options contracts based on the Standard & Poor's 500 (SPX) index, which comprises 500 of the largest publicly traded companies in the United States. The SPX index is widely regarded as a barometer of the overall health and performance of the US equity markets, making it a popular choice among traders and investors. Through the application of this strategy, we seek to demonstrate the practical utility and effectiveness of our proposed methodology within the realm of options trading and financial markets more broadly.

5.1.1 Evaluation Metrics

To measure the performance of both the baseline trading strategy and our proposed methodology, we utilize a set of standard metrics, including the premium capture rate, MAR ratio, and maximum drawdown.

Premium Capture Rate

The premium capture rate [PCR] is a fundamental metric used to assess the effectiveness of options trading strategies in capturing premiums. Options contracts are associated with premiums, representing the price paid by buyers to sellers for the right to buy or sell the underlying asset at a predetermined price within a specified period. The premium capture rate quantifies the proportion of these premiums that a trading strategy successfully captures or retains over a given period. A higher premium capture rate indicates greater efficiency in capitalizing on available premiums, signifying potential profitability. PCR above 0 means strategy is profitable.

Maximum Drawdown

The maximum drawdown [MDD] represents the maximum peak-to-trough decline in the value of a trading account or portfolio during a specific period. It is a crucial metric for assessing the downside risk and resilience of a trading strategy. A smaller maximum drawdown indicates lower volatility and a more stable performance trajectory, thereby enhancing investor confidence and mitigating the potential for significant capital erosion during adverse market conditions. Lower the drawdown the better.

MAR Ratio

The MAR ratio is a critical risk-adjusted performance metric utilized in financial analysis to evaluate the profitability of an investment strategy in relation to its risk exposure. Essentially, the MAR ratio quantifies the relationship between the average annual return generated by an investment and its maximum drawdown, which represents the largest peak-to-trough decline in value over a specific period. In essence, the MAR Ratio provides a measure of how much return an investment yields relative to the level of risk it assumes.

In practical terms, the MAR ratio is computed by dividing the average annual return of a trading strategy by its maximum drawdown. This calculation offers valuable insights into the risk-adjusted returns achieved per unit of risk undertaken by the strategy. A higher MAR ratio signifies superior risk-adjusted performance, indicating that the strategy delivers greater returns relative to the extent of its drawdowns. As such, the MAR ratio serves as a crucial benchmark for assessing the effectiveness and efficiency of a trading strategy, enabling investors to gauge its ability to generate profitable returns while effectively managing downside risk. As far as the number goes, higher the MAR the better, usually trader aim for MAR higher than 1, MAR higher than 2 is considered good.

5.1.2 Baseline strategy definition

In order to comprehensively evaluate the contribution of our proposed methodology to the field of trading strategies, it is imperative to establish a baseline against which its performance can be benchmarked. This section delineates the key evaluation metrics employed to rigorously assess the efficacy of our methodology within real-world trading environments. To facilitate a robust evaluation, we first define a baseline trading strategy. This strategy serves as a reference point for comparison, allowing us to gauge the relative performance and impact of our proposed methodology. In this subsection, we provide a detailed description of the trading strategy employed in our study. The strategy is designed to leverage 0 DTE (Days to Expiration) options contracts on the SPX index within the US equity markets.

- Each trading day, if a 0 DTE option chain is available, the following steps are executed:

- Sell short a 0 DTE put contract at the market opening (9:32 AM) with a premium closest to $100.

- Simultaneously, purchase a long 0 DTE put contract with the lowest available tick price, $5.

- Sell short a 0 DTE call contract at the market opening (9:32 AM) with a premium closest to $100.

- Simultaneously, purchase a long 0 DTE call contract with the lowest available tick price, $5.

- The exit criteria for the strategy is contingent upon the anticipated outcome where the options contracts expire worthless, thereby requiring no further exit action.

- Risk management protocols are integrated into the strategy to mitigate potential losses. If, at any point during the trading day, either the short call or short put option reaches a premium value of $300, both short positions are bought back to limit losses and preserve capital.

This strategy capitalizes on short-term price movements and intraday volatility within the US equity markets, leveraging the unique characteristics of 0 DTE options contracts to execute trades with rapidity and precision. Through the systematic application of this strategy, we aim to demonstrate its efficacy in generating consistent returns while effectively managing risk exposure.

5.1.3 Proposed strategy

In our pursuit to rigorously assess the efficacy of the proposed methodology, we endeavor to incorporate predictive analytics into our trading strategy. Specifically, we aim to utilize predictive models to forecast the daily volatility of the SPX index. Subsequently, we will establish a threshold th that serves as a determinant for the appropriateness of entering positions on a given day, contingent upon the relationship between our predicted volatility and the implied volatility derived from option prices [33].

The proposed strategy embodies a straightforward approach: upon extraction of option prices to initiate a trade, implied volatility is readily available. Our focus then shifts to predicting volatility, with our target variable being the difference between the high and low prices for the subsequent day. To achieve this, we will leverage the same dataset as delineated in the results section of the equity dataset, ensuring consistency and reliability in our analysis.

By integrating predictive analytics into our trading methodology, we aim to enhance decision-making processes by preemptively gauging market volatility dynamics. This proactive approach not only enables us to optimize entry timings (or lack of) but also affords us the opportunity to mitigate potential risks associated with heightened volatility levels. Through empirical validation and robust analysis, we endeavor to underscore the practical utility and effectiveness of our methodology in real-world trading scenarios.

5.1.4 Results

We conducted a comprehensive simulation of the trading strategy using the Option Omega backtesting software [39]. Given our prior experience in live trading with the same strategy, we have a high level of confidence in the comparability, if not equivalence, of the backtest results to live trading outcomes, particularly concerning key performance metrics. Our presentation of the results encompasses two distinct formats: a detailed tabular representation and a visual depiction of the profit curve.

The simulation spanned the period from January 1st, 2018, to January 1st, 2024, during which we traded the strategy on every available trading day featuring 0 DTE options. Consistent parameters were maintained for both the backtest and live trading scenarios, including an assumed commission fee of $1, entry slippage of $5, and exit slippage of $10. Additionally, we initiated trading with a starting capital of $100,000 and traded with a lot size of 1.

In the subsequent sections, we present the detailed outcomes of our simulation in a tabular format, cross-referenced as Table 5.1. Additionally, we provide a visual representation of the profit curve, offering a comprehensive overview of the strategy's performance dynamics over the specified timeframe. Through these comprehensive presentations, we aim to provide a thorough evaluation of the strategy's efficacy and performance characteristics, facilitating a nuanced understanding of its potential applications and implications in real-world trading contexts.

Strategy	PCR	MDD	MAR	Win %
Baseline Strategy	15.5%	-3.5%	1.3	55.1%
Proposed Enhanced Strategy	26.8%	-1.3%	3.6	61.4%

Table 5.1: Comparison of trading results between the baseline strategy and the proposed enhanced strategy.

We observe that the premium capture rate for baseline strategy is already relatively high at 15.5%, meaning a trader would get on average 15.5$ per 100$ risked. The results presented in the table are after commissions and simulated slippages. More importantly, we can see that proposed enhanced strategy has outstanding 26.8% PCR, representing a substantial improvement for an already profitable strategy. Simultaneously, the MAR ratio nearly triples to 3.6, indicating a significantly smoother and more consistent trading experience compared to the baseline scenario. This enhanced performance is also reflected in the improved win rate percentage. Figure 5.1 illustrates the profit curves for both strategies, allowing for a visual comparison of their performance.

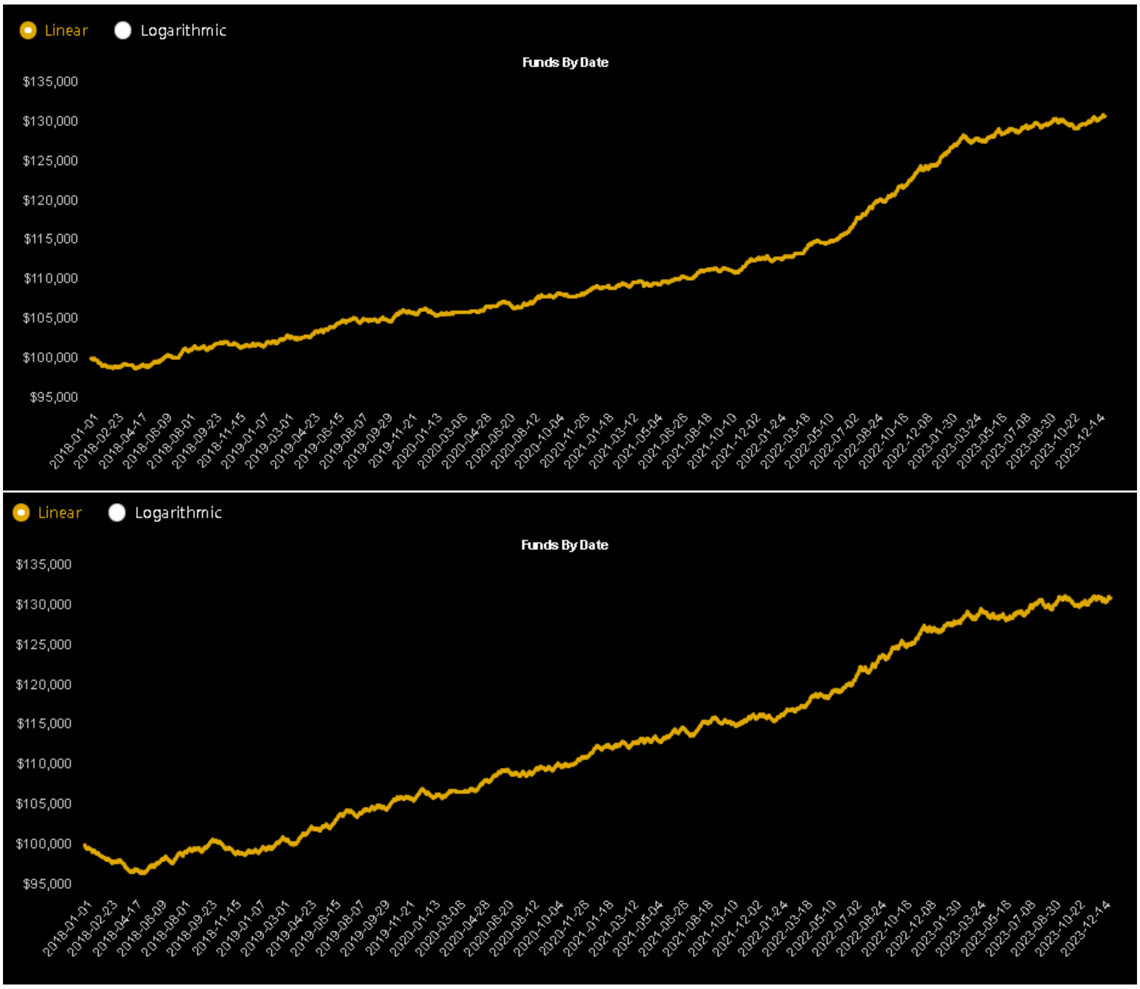

Figure 5.1: Profit curves for the proposed enhanced strategy (top) and the baseline strategy (bottom).

It is important to note that while the baseline strategy's nominal dollar returns are only slightly lower than those of the enhanced strategy, the baseline strategy involves significantly more trades, almost 50% more. This implies that, by employing the enhanced strategy, traders could allocate their capital elsewhere to achieve even greater returns or simply take more days off, resulting in a considerable reduction in the time spent monitoring the markets while earning the same income. This not only enhances capital efficiency but also improves the overall quality of life for traders by reducing the time commitment required for successful trading.

5.2 Predicting AI Skill Demand

The rapid evolution of technology has led to an exponential growth in demand for professionals skilled in AI, resulting in a significant talent gap in many countries. Analyzing the job demand in specific countries, such as the US, Germany [40], China [41], and India [42], offers valuable insights for policymakers, local businesses, and researchers. This understanding can help identify industrial hubs of technology, map the evolution of skills, and assess the effects on the global innovation ecosystem [43], [44].

The increasing demand for AI skills has been observed from 2010-2019 in the US economy across most industries and occupations, with the highest demand in IT occupations, followed by architecture and engineering, scientific, and management occupations [45]. A rapid increase in demand for machine learning (ML) skills has been documented since 2016, especially in the IT, finance, and professional services industries [42]. As AI transforms various industries, it is expected to create millions of new jobs by 2025, with many new roles emerging due to the collaboration of humans, machines, and algorithms [46].

The increasing use of artificial intelligence means that companies must integrate new technologies and also focus on continuous employee training. This approach helps ensure that workers stay skilled in using AI tools, keeping companies competitive and effective. Furthermore, it's important for policymakers to understand how AI affects jobs. This understanding is essential for developing education programs and policies that help build a workforce capable of adapting to the changes brought about by AI.

5.2.1 Identifying AI job postings

With the increasing demand for Artificial Intelligence (AI) talent in various industries, identifying and defining what constitutes an AI job posting has become a challenging task. Job postings that require AI skills and expertise can be ambiguous and heterogeneous, making it difficult to distinguish between them and non-AI related jobs. This is especially true for job postings in fields where AI skills are not necessarily required but can be beneficial. In this paper, we aim to provide a clear and consistent definition of AI job postings by analyzing a diverse set of job postings from 16 different countries. By defining what constitutes an AI job posting, we can better understand the trends and demand for AI talent, and provide insights into the skills and qualifications required for AI-related jobs.

1. **Filter for category**: At first, we utilize the Adzuna categorization to filter for IT-related job posting. It worth noting that this categorization is role-based not industry based. For example, a software engineering job in healthcare company is still considered IT. Hence by filter to IT category, we ensure the filtering for the relevant roles, while being cross-industry.

2. **Select a representative sample**: From the IT job postings, we select a random sample of 20 million job posting to perform the analysis.

3. **Semantic annotation**: For each selected job posting, we semantically annotate the description with Wikipedia concept, using wikifier [47]. wikification enables identifying related concepts from the job posting in different countries/languages and map them to the English Wikipedia concepts/articles.

4. **Selecting AI related job postings**: following the methodology used in [43], but adapting it to concepts instead of keywords. A job posting is considered AI if it's tagged at least 3 of the following concepts: "machine learning", "big data", "data science", "artificial intelligence", "deep learning","natural language processing", "tensorflow", "pytorch", "computer vision", "unsupervised learning", "supervised learning". The reason to of selecting at least 3 is to be on the conservative side and avoid job postings that just mention the names as buzz words or preferred to have, and focusing on job postings that involves AI for substantial part of the work.

5.2.2 Calculating AI relative job demand

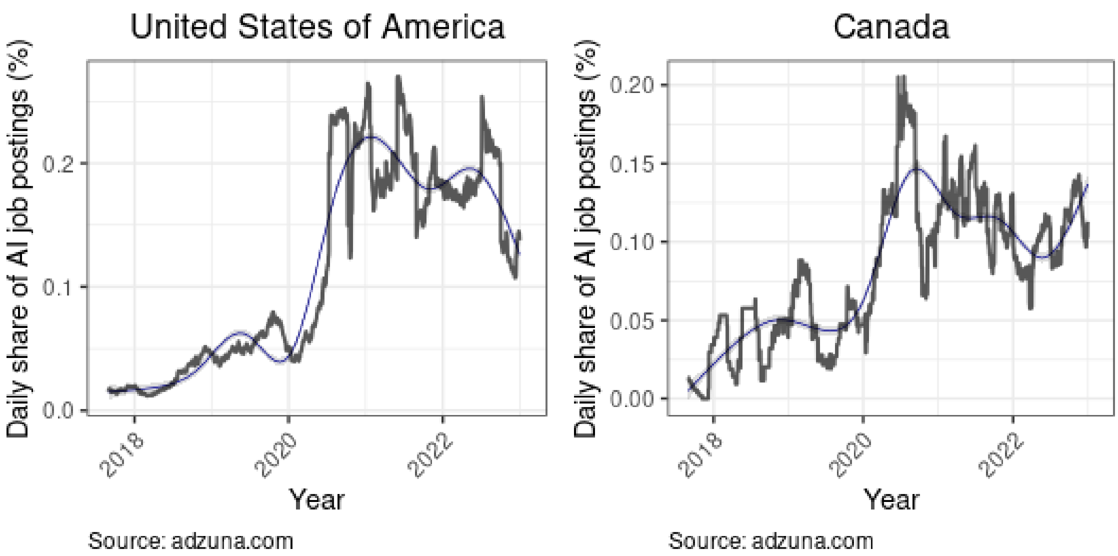

Figure 5.2: Comparison between relative AI demand between United states and Canada

To analyze the AI demand across countries, we calculate the AI demand timeseries by country. To do that we apply the following steps:

1. **Calculating daily AI and IT job posting by country**: we count the number of AI and IT job posting daily for each country

2. **Calculate AI relative demand**: We calculated the ratio of how many of IT jobs are AI related as percentage (1% means that 10 of 1000 IT jobs were AI).

3. **Apply smoothing**: using 90-days moving average

4. **Differentiate timeseries**: to make it stationary

5. **Apply z-transformation** to normalize timeseries

6. **Re-sampling timeseries** to quarter time steps with one week offset between each step

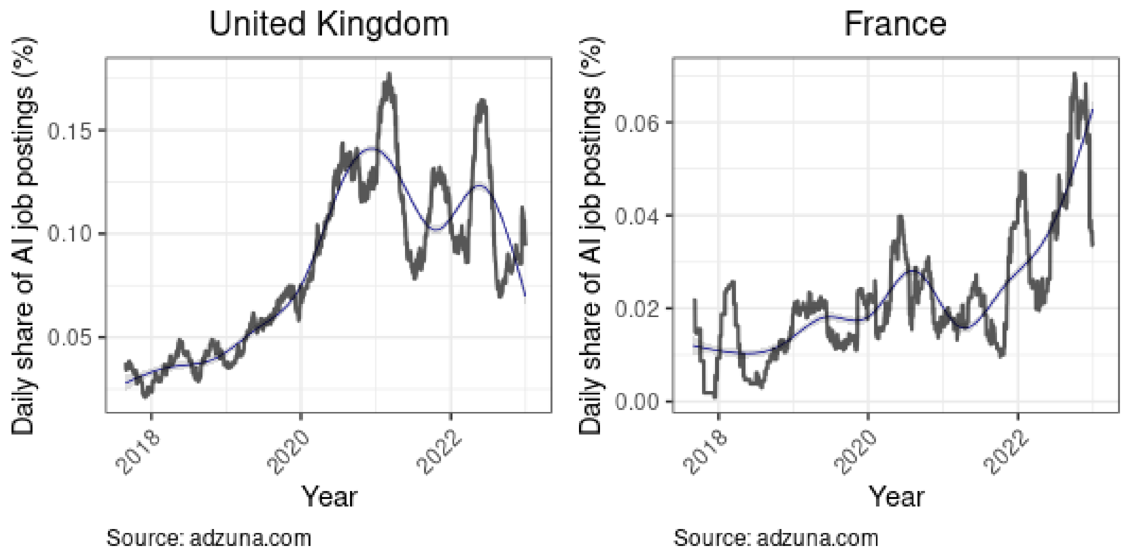

Figure 5.3: Comparison between relative AI demand between United Kingdom and France

Figure 5.2 displays the trend of AI demand over the years from 2018 to 2023. The graph indicates a significant increase in demand for AI in recent years, with a noticeable surge in 2020 due to the COVID-19 pandemic, followed by a normalization. The trend is projected to continue its upward trajectory, with a substantial spike observed in late 2022. The findings suggest that AI is rapidly becoming an integral part of various industries and sectors, with businesses seeking to leverage the technology's benefits to enhance efficiency, productivity, and profitability. Moreover, the rise in AI demand is expected to have significant implications for policymakers, who must ensure that the appropriate regulations and policies are put in place to govern AI's ethical use and prevent potential harm to society. Overall, the trends indicate that AI demand is set to increase significantly in the coming years, creating vast opportunities for individuals and organizations to develop innovative solutions that harness the full potential of AI technology.

Figure 5.3 shows the demand for AI in Europe, specifically in the United Kingdom and France. Both countries show a gradual increase in demand from 2018 to early 2023, with some minor fluctuations. The trend appears to be relatively flat compared to the sharp increase seen in the US. However, it is worth noting that the demand for AI in Europe may not necessarily reflect the actual investment and progress being made in the field, as there may be other factors at play such as regulatory challenges and cultural differences. Nevertheless, the steady increase in AI demand in Europe suggests that there is a growing interest and awareness of the potential benefits that AI can bring, and policymakers should take note of this trend and continue to support and invest in the field.

5.2.3 Skills shortage - understanding the demand

In this chapter, while the following subsection is not strictly relevant to the model predictions, we consider it important to discuss in order to understand the context of what we are modeling. One of the key advantages of the semantic annotation process is its capability to extract skills associated with job postings. By extracting and analyzing these concepts, we gain a deeper insight into the skills landscape specific to AI-related jobs. Furthermore, this process aids in identifying potential skill shortages within the AI job market across different countries. For the purposes of our analysis, we define 'skills' as concepts related

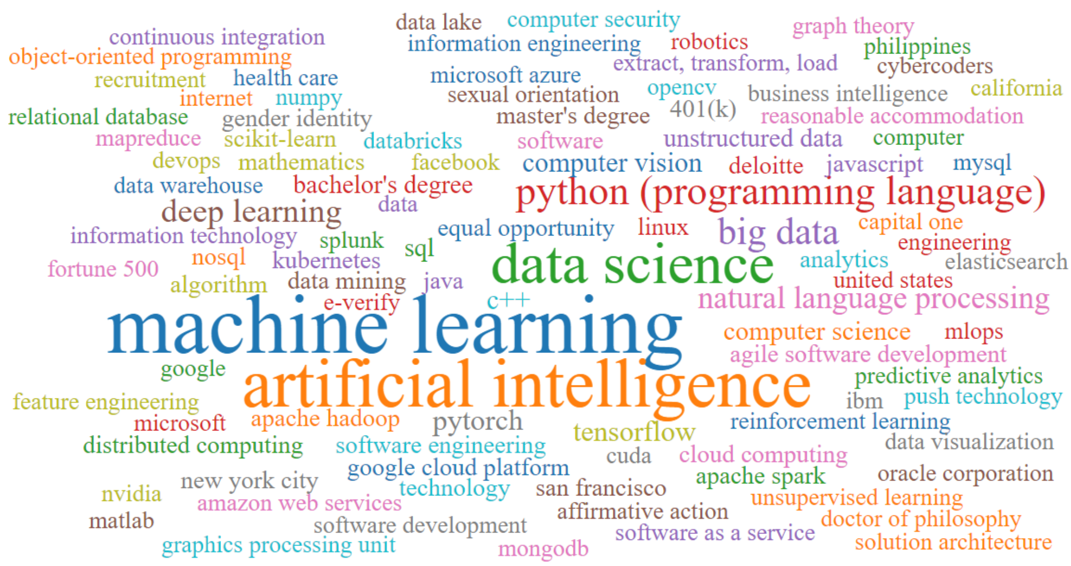

Figure 5.4: Top 100 skills by frequency.

to AI jobs that are extracted from job postings and aligned with corresponding entries in Wikipedia. This added context is vital for a comprehensive understanding of the dynamics at play in the AI employment sector, which in turn, enriches the foundational knowledge that supports our model predictions.

Figure 5.4 below shows the landscape of skills in jobs, highlighting the top 100 skills by relevance. The top concepts include the ones that match the filter, namely artificial intelligence, machine learning, and data science, but also others like python, and c++. Other concepts include skills and roles around data science, like data visualizations and data mining, and MLOps. Tools and databases like Apache Hadoop, Kubernetes, MongoDB, and Elasticsearch, indicates the need for having engineering skills for doing AI in the industry. Career-related skills like higher education and academic degrees highlight the yet required higher educational degrees to do most of the AI and machine learning work.

While the skills landscape provides meaningful but obvious results, a more useful analysis of skills is to identify for each country, what are the most needed skills required, as this helps policymakers develop programs for handling shortages. The intuition behind identifying skill shortages is the inverse correlation between jobs in demand and skills supply (i.e. supply and demand). That is, if a skill is associated with a lot of job postings that take a long to fill, that means the skill is in demand but there isn't enough supply of people to fill it, hence a skill shortage.

More concretely, we utilize a process that involves observing the duration of job postings on Adzuna. First, We define J_i as the set of job postings that closed within the i_{th} week of its posting. Second, we identify two points of interest; $J_{t0} = J_1$ as being the set of in-demand jobs and $J_t = J_9$ as being the set of jobs filled within the 9th week. Due to the distribution of jobs, $J_9 \approx \sum_{i>=9} J_i$, hence it represents the jobs that are in demand but not filled for more than two months approximately due to not enough supply.

Now for these two timestamps, we count the distribution of the skills, to calculate the decline in demand for each skill, we can define a skill shortage if its decline (first derivative) is larger than the average of all skills, i.e. takes longer than average to get filled. More concretely:

Then we define the skill decline $D_s$as follows

$$D_s = \frac{\sum_{\forall j \in J_t} P_{j,s} - \sum_{\forall j \in J_{t0}} P_{j,s}}{t - t0} \tag{5.1}$$

Where $P_{j,s}$ is 1 if job posting j is tagged with skill s, and 0 otherwise. We calculate the Demand score by the country for each of the 100 most common concepts. The demand allows us to gain insights into the skill sets that are sought after by employers (high in-demand score) and the ones that are not (low on-demand score). A high-demand skill will have more occurrences in in-demand jobs than in not-in-demand ones. Then the skill shortage score for each skill is defined as the demand score of the skill divided by the average demand of the skills minus one:

$$SS_s = \frac{D_s}{\frac{1}{|S|}\sum_{s \in S} D_s} - 1 \tag{5.2}$$

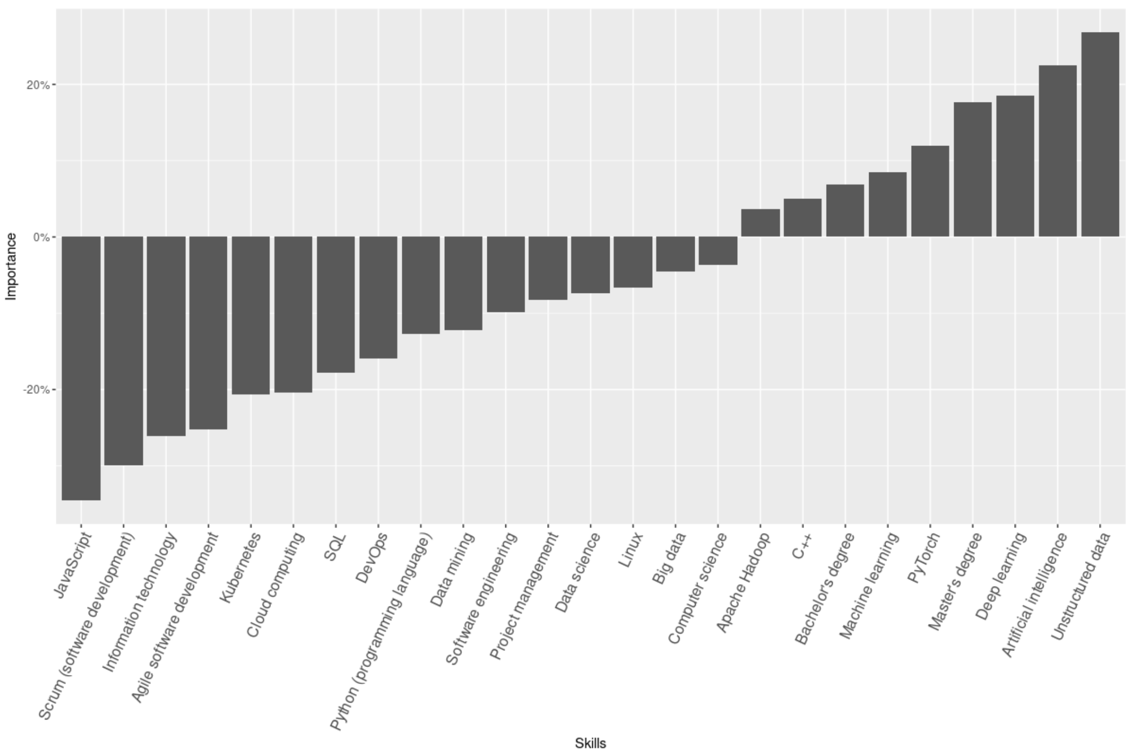

Figure 5.5: Skill shortage of US job postings

A positive score (ratio above 1) indicates that jobs with this skill remained online longer than average, indicating a shortage of this skill. Conversely, a negative score (ratios below 1) indicates that jobs with this skill were filled quickly, suggesting an abundance of that skill. Looking at the shortage of skills in the US in Figure 5.5, we observe that skills with the most shortage (the right most skills in the barchart) are "Unstructured data", "Artificial intelligence", "Deep learning", "Master's degree", whereas some of the top skills with no shortage of supply (the left most skills in the barchart) are "JavaScript", "Scrum and agile development", "Information technologies", "Kubernetes".

5.2.4 Forecasting of AI demand

Following an examination of artificial intelligence (AI) demand across seven countries, our attention shifted to modeling this demand using diverse methods. Our analysis spanned from 2022 onwards for test data, covering a one-year period. To prevent data leakage during the smoothing step, we excluded timesteps that only partially overlapped with 2022 from the training data. Our primary focus was on predicting AI demand for the upcoming quarter—a crucial aspect for businesses and policymakers in planning resource allocation, staffing, and budgeting. We employed both univariate and multivariate models, with detailed results provided below.

5.2.4.1 Univariant predictions

In our univariate modeling analysis, we compared four different models: a baseline model (which uses the last known value), a generalized linear model (GLM), a random forest model, and a univariate long short-term memory (LSTM) model. The LSTM model used an architecture of three stacked LSTM layers with 32 units each, utilizing the last 6 steps of data to predict the next step. Optimization was performed using the proposed method.

Table 5.2.4.1 below presents the mean squared error (MSE) for each model across different countries. The baseline model, which predicts demand based on the mean value for each country, showed MSEs ranging from 1.54 in India to 3.292 in the US. The GLM, random forest, and univariate LSTM models all outperformed the baseline, with MSEs as low as 0.709 for random forest in Singapore and 1.768 for LSTM in Canada. These results indicate that these models can enhance the accuracy of AI demand forecasting, with each model performing best in different countries. The following sections will further examine each model's results and their implications for forecasting AI demand.

	US	CA	UK	FR	IN	SG	AU	ALL
Baseline (last value)	3.292	2.770	2.173	2.196	1.540	2.378	1.691	2.291
GLM	1.378	**1.503**	1.217	1.228	1.254	1.468	0.978	1.289
Random forest	**1.310**	1.534	**1.129**	1.102	**1.047**	1.461	**0.709**	**1.185**
LSTM	1.631	1.768	1.285	**1.095**	1.193	**1.348**	0.811	1.304

Table 5.2: Comparison of Mean Squared Errors for Univariate AI Demand Models Across Seven Countries.

In more detail, we found that the GLM and random forest models generally outperformed the LSTM model in most countries. The GLM model achieved the lowest MSE in four out of seven countries, while the random forest model performed best in two countries, including Singapore. However, the LSTM model outperformed both the GLM and random forest models in France, suggesting that it may be more effective in regions with less predictable demand patterns.

5.2.4.2 Multivariate Predictions

Following the promising results from univariate modeling of AI demand, we now examine the effectiveness of multivariate modeling using the LSTM model with the same input data. In this section, we compare the results from the multivariate LSTM model to those from the previous univariate LSTM model. The table presents the mean squared errors (MSE) for the multivariate LSTM model across all seven countries. Notably, the multivariate LSTM model achieves lower MSEs than the univariate model in five out of seven countries and shows an overall improvement in prediction accuracy. These findings suggest that incorporating multiple variables into the LSTM model enhances its predictive accuracy for AI demand, even for short-term forecasts one quarter ahead.

	US	CA	UK	FR	IN	SG	AU	ALL
Univariate	**1.631**	1.768	1.285	1.095	1.193	1.348	0.811	1.304
Multivariate	1.659	**1.649**	**1.083**	**1.062**	**1.074**	**1.318**	**0.703**	**1.162**

Table 5.3: Comparison of mean squared errors between univariate and multivariate LSTM model predictions across seven countries.

The key finding is the performance of multivariate prediction. Upon detailed examination, our method either surpasses or significantly narrows the performance gap compared to the top performer in the univariate prediction test. The reason our approach does not consistently outperform across all datasets in Table 5.2.4.2, as it does with financial data from both Section 4.3 and 5.1, can be attributed to the level of stochastic variability required in the data for our method to perform optimally. This variability is essential for our predictive model to fully leverage its capabilities, suggesting that the method's effectiveness is partly dependent on the specific characteristics of the dataset.

Chapter 6

Conclusions

SC1 - Formal definition and development of the proposed method.

Aim: Formal definition and development of a novel optimization technique based on smart denoising specialized for recurrent neural networks.

In Chapters 2 and 3, we introduced a new denoising optimization methodology along with a comprehensive evaluation framework. The proposed methodology consists of three key components. First, we apply empirical normalization to all central moments of the normal distribution, an essential albeit expected aspect of the method. Next is latent optimization, which ensures that the model accurately represents the underlying space, even within the intermediate layers of the network. Lastly, we implement a gradual reduction of noise addition throughout the learning process, aligning more closely with the true underlying function as the model progresses.

SC2 - Evaluation of the proposed method on financial data.

Aim: Evaluation of the proposed method using real-world financial datasets. These datasets represent a complex, unknown stochastic process and include data from thousands of equity prices spanning over a twenty year period.

We have utilized a comprehensive dataset comprising over 2000 equities spanning the last 20 years. Through rigorous statistical analysis, we demonstrated the superior performance of our proposed method across five distinct hypotheses. We delved into the exploration of the parameter space, seeking to understand the reasons behind the method's effectiveness, which we attribute to the high degree of stochasticity present in the data. Additionally, we illustrated how our methodology could enhance an already state-of-the-art trading strategy, pushing the boundaries of its performance. The evaluation of the strategy was meticulously conducted using risk-adjusted metrics, and the outcomes were clearly substantiated through detailed profit and loss/risk adjustment analysis. This evidence underscores the robustness and potential of the proposed method in generating superior trading insights and results.

SC3 - Evaluation of the proposed method on text data.

Aim: Evaluation of the proposed method using real-world example of text data, specifically skill demand datasets derived from multilingual job postings over sixteen countries spanning over five year period.

In the thesis, we have modeled skill demand by extracting insights from textual data across 16 diverse countries. Our analysis confirmed the outperformance of the proposed method against five distinct hypotheses, demonstrating its robust applicability. Furthermore, we embarked on an exploration of the parameter space to identify instances where the method underperformed, particularly when compared to its efficacy with financial data. This exploration was crucial for highlighting potential limitations and refining our approach. A practical example of our methodology was illustrated through the modeling of AI skill demand in multiple countries, showcasing its real-world applicability. The study also outlined ways in which the baseline models could be enhanced using our proposed method, suggesting paths for future improvements and adaptations in the field of skill demand analysis.

6.1 Future Work

This thesis has established a basic understanding of how our method's performance varies when applied to different data types characterized by unique stochastic properties. Although the study covers several scenarios, there remains much to explore in future work.

A crucial area for further investigation is how the type of stochastic processes underlying different datasets affects the performance of our method. Financial data often follows a pattern known as Brownian Motion, which differs from patterns seen in other types of datasets, like job postings. Future research should look into how our optimization method performs across various types of stochastic data. Understanding these relationships will improve our ability to tailor optimization techniques to specific data characteristics.

Additionally, finding the optimal level of noise tolerance for a dataset is an important next step. Initially, this research should focus on the financial sector, where stochastic data is commonly found. It's vital to determine how adjusting the noise level affects the method's ability to make accurate predictions and how these predictions can improve trading strategies. This involves looking at the effects on risk-adjusted performance measures, such as the Sharpe and MAR ratios.

Another important area for future research is to identify when our method does not perform well. Our initial observations indicate that too much noise in unsuitable datasets can reduce the model's effectiveness on new data. Future work should extend these empirical tests and also develop theoretical models to better understand why and when the method might fail. This will help in making the method more reliable and accurate.

By exploring these areas, future research can greatly enhance the application of stochastic optimization methods across different fields, ensuring they are both effective and adaptable to the complexities of real-world data.

References

[1] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[3] M. A. Kramer, “Nonlinear principal component analysis using autoassociative neural networks,” *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[4] F. Black and M. Scholes, “The pricing of options and corporate liabilities,” in *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 1: Foundations of CCA and Equity Valuation*, World Scientific, 2019, pp. 3–21.

[5] K. R. Sircar, G. C. Papanicolaou, *et al.*, “Stochastic volatility, smile & asymptotics,” *Applied Mathematical Finance*, vol. 6, pp. 107–145, 1999.

[6] G. E. Uhlenbeck and L. S. Ornstein, “On the theory of the brownian motion,” *Physical review*, vol. 36, no. 5, p. 823, 1930.

[7] D. B. Madan, “Stochastic processes in finance,” *Annu. Rev. Financ. Econ.*, vol. 2, no. 1, pp. 277–314, 2010.

[8] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, “Deep learning for time series classification: A review,” *Data mining and knowledge discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[9] F. Sowell, “Modeling long-run behavior with the fractional arima model,” *Journal of monetary economics*, vol. 29, no. 2, pp. 277–302, 1992.

[10] G. E. Box and D. A. Pierce, “Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,” *Journal of the American statistical Association*, vol. 65, no. 332, pp. 1509–1526, 1970.

[11] W. Xu, Z. Fu, H. Li, J. Huang, W. Xu, and Y. Luo, “A study of the impact of covid-19 on the chinese stock market based on a new textual multiple arma model,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 16, no. 1, pp. 5–15, 2023.

[12] S. Mukherjee, B. Sadhukhan, N. Sarkar, D. Roy, and S. De, “Stock market prediction using deep learning algorithms,” *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 82–94, 2023.

[13] Q. Ding, S. Wu, H. Sun, J. Guo, and J. Guo, “Hierarchical multi-scale gaussian transformer for stock movement prediction.,” in *IJCAI*, 2020, pp. 4640–4646.

[14] J. Jelenčič and D. Mladenić, “Improving modeling of stochastic processes by smart denoising,” *Informatica*, vol. 46, no. 1, 2022.

[15] J. Jelencic and D. Mladenic, “Modeling stochastic processes by simultaneous optimization of latent representation and target variable,” *Proceedings of SiKDD*, 2020.

[16] J. Jelenčič, M. B. Massri, L. Todorovski, M. Grobelnik, and D. Mladenić, "Improving stochastic models by smart denoising and latent representation optimization [under minor revision]," *Information Sciences*, vol. tbd, tbd, 2024.

[17] N. Wax, *Selected papers on noise and stochastic processes*. Courier Corporation, 2014.

[18] B. W. Turnbull, "The empirical distribution function with arbitrarily grouped, censored and truncated data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 38, no. 3, pp. 290–295, 1976.

[19] H.-C. Chen and Y. Asau, "On generating random variates from an empirical distribution," *AIIE Transactions*, vol. 6, no. 2, pp. 163–166, 1974.

[20] P. Jaworski, F. Durante, W. K. Hardle, and T. Rychlik, *Copula theory and its applications*. Springer, 2010, vol. 198.

[21] H. Joe, *Dependence Modeling with Copulas*. CRC Press, 2014.

[22] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," *Advances in neural information processing systems*, vol. 29, pp. 901–909, 2016.

[23] M. Liu, W. Wu, Z. Gu, Z. Yu, F. Qi, and Y. Li, "Deep learning based on batch normalization for p300 signal detection," *Neurocomputing*, vol. 275, pp. 288–297, 2018.

[24] K. Y. Levy, "The power of normalization: Faster evasion of saddle points," *arXiv preprint arXiv:1611.04831*, 2016.

[25] J. Kolen and J. Pollack, "Back propagation is sensitive to initial conditions," *Advances in neural information processing systems*, vol. 3, 1990.

[26] R. Vidal, J. Bruna, R. Giryes, and S. Soatto, "Mathematics of deep learning," *arXiv preprint arXiv:1712.04741*, 2017.

[27] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[28] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[29] D. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*. 2014, `https://arxiv.org/abs/1412.6980`.

[30] B. Rosner, R. J. Glynn, and M.-L. T. Lee, "The wilcoxon signed rank test for paired comparisons of clustered data," *Biometrics*, vol. 62, no. 1, pp. 185–192, 2006.

[31] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.

[32] S. Taheri and G. Hesamian, "A generalization of the wilcoxon signed-rank test and its applications," *Statistical Papers*, vol. 54, no. 2, pp. 457–470, 2013.

[33] R. C. Merton, "Option pricing when underlying stock returns are discontinuous," *Journal of financial economics*, vol. 3, no. 1-2, pp. 125–144, 1976.

[34] *Yahoo Finance*, `https://finance.yahoo.com/`.

[35] J. J. Murphy, *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications* (New York Institute of Finance Series). New York Institute of Finance, 1999, ISBN: 9780735200661. [Online]. Available: https://books.google.com/books?id=5zhXEqdr%5C_IcC.

[36] *Adzuna*, https://www.adzuna.com/.

[37] J. Brank, G. Leban, and M. Grobelnik, "Annotating documents with relevant wikipedia concepts," *Proceedings of SiKDD*, 2017.

[38] J. Jelenčič, M. B. Massri, M. Grobelnik, and D. Mladenić, "A multilingual approach to analyzing talent demand in a specific domain: Insights from global perspectives on artificial intelligence talent demand," *IEEE Access*, 2024.

[39] *Option Omega*, https://docs.optionomega.com/.

[40] F. Stephany, *There is Not One But Many AI: A Network Perspective on Regional Demand in AI Skills*, en-us, Mar. 2020. DOI: 10.31219/osf.io/32qws. [Online]. Available: https://osf.io/32qws/ (visited on 03/24/2023).

[41] C. Sun and Y. Lu, "Prediction of Information Talent Demand Based on the Grayscale Prediction Model and the BP Neural Network," en, *Mobile Information Systems*, vol. 2022, e4050502, Aug. 2022, Publisher: Hindawi, ISSN: 1574-017X. DOI: 10.1155/2022/4050502. [Online]. Available: https://www.hindawi.com/journals/misy/2022/4050502/ (visited on 03/24/2023).

[42] K. Stapleton, A. Copestake, and A. Pople, "AI, firms and wages: Evidence from India," en, *SSRN Electronic Journal*, 2021, ISSN: 1556-5068. DOI: 10.2139/ssrn.3957858. [Online]. Available: https://www.ssrn.com/abstract=3957858 (visited on 03/24/2023).

[43] T. Babina, A. Fedyk, A. X. He, and J. Hodson, *Artificial Intelligence, Firm Growth, and Product Innovation*, en, SSRN Scholarly Paper, Rochester, NY, May 2022. DOI: 10.2139/ssrn.3651052. [Online]. Available: https://papers.ssrn.com/abstract=3651052 (visited on 02/27/2023).

[44] J. Klinger, J. Mateos-Garcia, and K. Stathoulopoulos, "Deep learning, deep change? Mapping the evolution and geography of a general purpose technology," en, *Scientometrics*, vol. 126, no. 7, pp. 5589–5621, Jul. 2021, ISSN: 1588-2861. DOI: 10.1007/s11192-021-03936-9. [Online]. Available: https://doi.org/10.1007/s11192-021-03936-9 (visited on 03/23/2023).

[45] L. Alekseeva, J. Azar, M. Gine, S. Samila, and B. Taska, *The Demand for AI Skills in the Labor Market*, en, SSRN Scholarly Paper, Rochester, NY, Oct. 2019. DOI: 10.2139/ssrn.3470610. [Online]. Available: https://papers.ssrn.com/abstract=3470610 (visited on 03/24/2023).

[46] M. K. Khan, "AI-enabled transformations in telecommunications industry," en, *Telecommunication Systems*, vol. 82, no. 1, pp. 1–2, Jan. 2023, ISSN: 1018-4864, 1572-9451. DOI: 10.1007/s11235-022-00989-w. [Online]. Available: https://link.springer.com/10.1007/s11235-022-00989-w (visited on 03/24/2023).

[47] J. Brank, G. Leban, and M. Grobelnik, "Annotating documents with relevant Wikipedia concepts," 2017. (visited on 03/01/2023).

Bibliography

Publications Related to the Thesis

Journal Articles

J. Jelenčič and D. Mladenić, "Improving modeling of stochastic processes by smart denoising," *Informatica*, vol. 46, no. 1, 2022.

J. Jelenčič, M. B. Massri, M. Grobelnik, and D. Mladenić, "A multilingual approach to analyzing talent demand in a specific domain: Insights from global perspectives on artificial intelligence talent demand," *IEEE Access*, 2024.

Conference Paper

J. Jelencic and D. Mladenic, "Modeling stochastic processes by simultaneous optimization of latent representation and target variable," *Proceedings of SiKDD*, 2020.

Publications Unrelated to the Thesis

Journal Articles

J. Jelenčič, M. B. Massri, L. Todorovski, M. Grobelnik, and D. Mladenić, "Improving stochastic models by smart denoising and latent representation optimization [under minor revision]," *Information Sciences*, vol. tbd, tbd, 2024.

Conference Paper

J. Jelenčič, J. M. Rožanec, and D. Mladenić, "Kl-adwin: Enhanced concept drift detection over multiple time windows," in *Central European Conference on Information and Intelligent Systems*, Faculty of Organization and Informatics Varazdin, 2022, pp. 49–54.

J. Jelenčič, "Modeling text data over time-example on job postings," in *Companion Proceedings of the Web Conference 2021*, 2021, pp. 721–722.

J. Jelenčič, "Predicting bitcoin trend change using tweets," *Proceedings of SiKDD*, 2019.

Biography

The author of this thesis is an emerging data scientist and researcher, originally from Ljubljana, with a solid background in financial mathematics. After completing both a bachelor's and master's degree in Financial Mathematics, with an emphasis on probability and stochastic processes, he joined the Jožef Stefan Institute (JSI) in 2019. Here, he has been actively involved in research within the realms of AI and machine learning.

His professional experience spans roles at Abanka, Triglav, and the OECD, where he supported data system management and contributed to projects focused on risk management. He enjoys exploring different trading strategies. In his own time, he engages in trading, applying theoretical insights to practical scenarios. This combination of academic dedication and practical application underpins his growing influence in the field of data science and financial analytics.