Predicting Dark Matter Presence Using Supervised Machine Learning

by

Examination Roll: 232107

A Project Report submitted to the
Institute of Information Technology
in partial fulfillment of the requirements for the degree of
Professional Masters in Information Technology

Supervisor: Professor Dr. Risala Tasin Khan

Institute of Information Technology
Jahangirnagar University
Savar, Dhaka-1342

September 2025

DECLARATION

I hereby declare that this thesis is based on my original research and findings. All sources and contributions from other researchers have been properly cited. This work has not been previously submitted, either in full or in part, for the award of any degree or qualification.

.

Roll: 232107

CERTIFICATE

The project titled "Predicting Dark Matter Presence in Galaxies Using Supervised Machine Learning" submitted by Student - S. M. JAHANGIR , ID: 232107, Session: Summer 23, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Professional Masters in Information Technology on the 11th of October 2025

Professor Dr. Risala Tasin Khan
Supervisor

BOARD OF EXAMINERS

Professor Dr. M. Shamim Kaiser
Professor, IIT, JU
PMIT Coordination Committee

Professor Dr. Risala Tasin Khan
Professor, IIT, JU
Member, PMIT Coordination Committee
& Director, IIT

Professor Dr. Jesmin Akhter
Associate Professor, IIT, JU
Member
PMIT Coordination Committee

Professor K M Akkas Ali
Associate Professor, IIT, JU
Member,
PMIT Coordination Committee

Dr. Rashed Mazumder
Associate Professor, IIT, JU
Member
PMIT Coordination Committee

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to everyone who played a role in the successful completion of this thesis. I am especially thankful to Dr. Risala Tasin Khan, a distinguished Professor at the Institute of Information Technology (IIT), Jahangirnagar University, for her invaluable guidance, thoughtful suggestions, and constant encouragement throughout this journey. Her generous provision of a rich collection of books, journals, and research materials greatly enhanced the quality of my work.

I am also deeply grateful to Dr. Risala Tasin Khan, Director of IIT, for her insightful advice and unwavering support. My sincere thanks go to all the esteemed faculty members of IIT for their helpful direction and encouragement. I also acknowledge and appreciate the contributions of those who, directly or indirectly, supported the successful completion of this research.

Finally, I would like to thank the administrative and support staff of IIT, Jahangirnagar University, as well as my friends, for their continuous encouragement and involvement throughout this academic journey.

ABSTRACT

Dark matter, accounting for a large part of the mass in the universe, yet elusive of detection in conventional observations because it does not interact with electromagnetic radiation. Prediction of the properties and dynamics of dark matter is hence a burning issue in astrophysics. In the following, we investigate the use of machine-learning (ML) approaches to predict dark matter properties from cosmological data and theory. Using many different supervised and unsupervised learning algorithms, such as decision trees, support vector machines or deep neural networks, we study a variety of input features including galaxy rotation curves, weak lensing data and cosmic microwave background measurements. Our results show that machine learning models can be used to infer patterns in large datasets which are the signature of dark matter distribution and characteristics, helping to shed light onto its elusive nature. Moreover, our predictive models give encouragingly accurate predictions for the dark matter density profiles and possible interactions with visible matter. This study highlights the power of ML in advancing the fundamental understanding of dark matter, serving as a harbinger for further research and experimental verification.

Keywords: Dark matter, machine learning, cosmological data, galaxy rotation curves, gravitational lensing, cosmic microwave background, prediction models, supervised learning, unsupervised learning, neural networks, astrophysics, density profiles, dark matter interactions, predictive modeling.

LIST OF ABBREVIATIONS

SMV	Support Vector machine
KNN	K-Nearest Neighbors
LR	Logistic Regression
ROC	The Receiver Operating Carve
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
RF	Random Forest
CMB	Cosmic Microwave Background
PCA	Principal Component Analysis
WIMP	Weakly Interacting Massive Particle
CDM	Cold Dark Matter
CV	Cross-Validation

LIST OF FIGURES

Figure

LIST OF TABLES

Table

TABLE OF CONTENTS

CHAPTER I

Introduction

1.1 Overview

Dark matter is among the most fascinating and enigmatic features of the universe. Although representing roughly 27% of the mass-energy content in the universe, this composition has yet to be detected directly.[2] Unlike normal matter, which feels electromagnetic forces, dark matter does not emit, absorb, or reflect light so it cannot be viewed using telescopes. Attendant necessarily to such a particle of "dark" matter would be the creation of theories to account for all the rest of what we observe without it. Its existence was first proposed in the early 20th century, as it had become apparent that visible material alone could not explain all the observed motion of galaxies and galaxy clusters.

Dark matter is detected and analyzed indirectly through gravitational effect on visible matter, which includes its influence upon the motion of galaxies and stars, accelerating the universe's flow in space-time (also cutting forth its mass content), acting as a bendable mass source for gravitation lensing, and merging corpuscular small-scale elements into large-scale structures. One of the most potent evidence for dark matter arises from the fact that galaxies spin far more rapidly than can be explained by their visible content alone.[3] This indicates an immense amount of invisible mass is exerting extra gravitational force, and thereby preventing galaxies from flying apart.

Dark matter has not yet been directly observed, but its composition is assumed to be the same that of matter that does absorb radiation like light. The more promising sought-after dark matter particles are WIMPs, axions, or sterile neutrinos, none of which have been observed experimentally yet.

Astrophysics and cosmology have made dark matter a very hot topic in recent years with many theories and models being proposed. Scientists want to learn about

its properties, where it is located and how it interacts with “ordinary” visible matter, which could have broad implications regarding the structure and evolution of everything in the universe. Machine learning (ML) methods have lately shown the potential to effectively process multiple sources of cosmological data, recognize patterns, and make inference about the behavior and properties of DM in a manner that opens up new paths for investigation of this elusive entity.

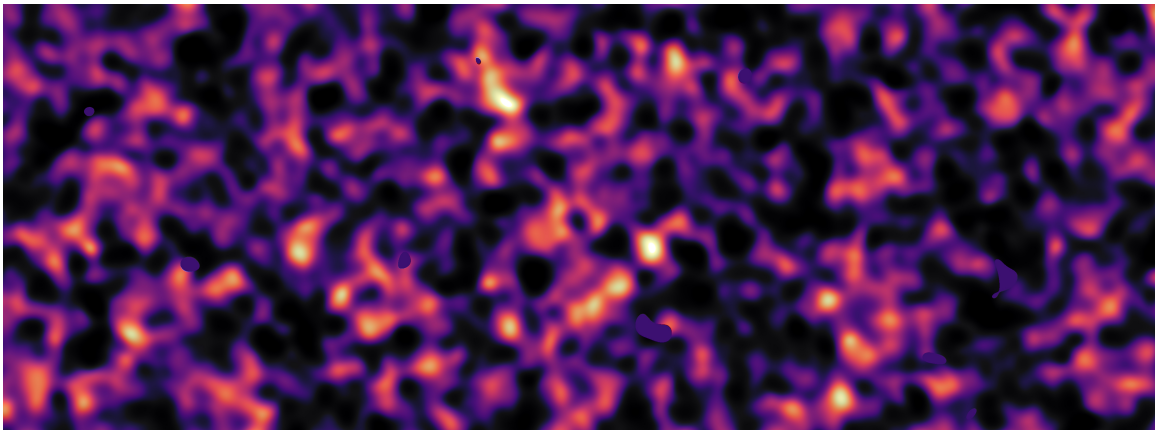

Figure 1.1: Dark matter map for a patch of sky based on gravitational lensing analysis of a Kilo-Degree Survey [1]

1.1.1 Observational Tests of the Dark Matter Scenario

Despite being invisible, dark matter has a substantial gravitational pull on visible matter and therefore numerous astrophysical observations have helped establish the existence of dark matter. A key early indication comes from observing the rotation curves of galaxies. According to Newtonian mechanics, stars and gas orbiting in a galaxy should slow down the farther they get from its center, where less mass is concentrated. But when we turn to observations of spiral galaxies, we find that stars at their outskirts are spinning, but they’re spinning far too quickly considering all the mass we can account for by the light produced from them. That implies the existence of more, unseen mass tugging on it gravitationally, a form of matter known as dark matter. The existence of this unseen mass, commonly known as the "dark halo," is required to account for both the observed flat rotation curves and the differences between predicted and actual stellar velocities within galaxies.

The Legacy of Gravitational Lensing is also a vital line of evidence for general relativity, the gravitational bending of light that Einstein’s 100-year-old theory predicts. When a beam of light from a very distant object such as a galaxy passes close to something big, like another galaxy cluster (or other matter of lesser mass), the

path of that light beam gets bent and distorted, and you get warped looking images of background objects. The magnitude of this effect is directly related to the mass of the lensing body. Measurements of gravitational lensing around galaxy clusters provide evidence that the mass indicated by the lensing is much larger than that comprised only of visible matter contained in such clusters, implying that dark matter exists. One of the best known cases, the Bullet Cluster, has a separation between visible (X-ray) and dark matter (inferred from lensing), providing even more striking evidence for dark matter in galaxy clusters.[4]

Evidence for dark matter The Big Bang left behind the Cosmic Microwave Background (CMB). These differences are those of temperature fluctuations of the CMB, which in turn reflect small density variations in the matter content and distribution of the early universe. These fluctuations are what enables cosmologists to infer the distribution of visible and invisible matter at that time. The results from the CMB, in particular the angular scale subtended by the first acoustic peak, provide extensive evidence for dark matter. Dark matter was essential for the gravitationally driven growth of cosmic structure that is encoded in the fluctuations in the CMB. By observing these fluctuations, researchers have measured the quantity of dark matter in the universe and verified its vital role in defining the cosmos.

Taken together, these three observational facts— galaxy rotation curves, gravitational lensing and Cosmic Microwave Background are a compelling suite of evidence for the reality of dark matter. Not only do they show that dark matter is critical for the creation of galaxies and galaxy clusters they also offer clues on what role it played in the early universe.

1.1.2 Early Dark Matter Theories

It was around the early 20th century that scientists first began to realize that there were discrepancies between the motion of cosmic objects as they observed it and how much visible matter they could count. First evidence of what would eventually become known as dark matter comes from the research of the rotation curves of galaxies and member galaxy motions in clusters.

In 1933, the Swiss astronomer Fritz Zwicky observed one of the first pieces of evidence with dark matter. While observing the Coma galaxy cluster, Zwicky realized that galaxies in a cluster were moving at much higher velocities than could be explained by the matter we could see.[5] In terms of classical Newtonian mechanics, the gravitational attraction of the visible matter in the cluster should have decelerated the motions of galaxies and dispersed them until they could eventually escape from

the cluster. But Zwicky's measurements revealed that the total mass of the cluster didn't suffice to explain what was keeping the galaxies tied up. He posited that there was a mysterious force of otherwise undetectable "missing mass" which was just sufficient to bind the galaxies. This "missing mass" would foreshadow the modern idea of dark matter. Zwicky even went so far as to call the substance "dunkle materie" or dark matter, although the idea went mostly unheralded at the time.[5]

The idea received a boost in the 1970s after astronomer Vera Rubin, building on Zwicky's research, analyzed the rotation curves of galaxies.[3] Rubin and her colleague Kent Ford discovered that stars on the outskirts of spiral galaxies were moving at unexpectedly high rates, given only the amount of visible matter in these galaxies. Also like Zwicky, Rubin determined that some type of dark or hidden mass must be exerting gravitational forces upon the stars in the arms, reinforcing her belief in the existence of dark matter. Her work was instrumental in placing dark matter at the center of modern cosmology.

When the notion of dark matter was first proposed, it originated from our inability to explain the motions of galaxies (as well as galaxy clusters) using gravity and regular baryonic matter alone, and scientists were free to imagine what kind of form this mysterious missing mass could take. It was speculated by some that it could be made up of ordinary, invisible or "dark" matter like faint stars or clouds of gas (so-called baryonic matter). But None of these is consonant with all the observed facts, notably the large structure of universe and detailed dynamics of galaxy clusters.

As dark matter was better understood, it turned out that it couldn't just be regular old matter. This, in turn, eventually gave rise to the more sophisticated theories that exist today which postulate cosmological dark matter is primarily made of non-luminous, non-interacting particles. These particles (which interact via gravitational and, potentially, the weak force) give rise to more distinctive dark matter theories such as WIMPs and axions that will be discussed later in the section.

So, that's an overview of the early theories of dark matter which were developed mostly in response to astronomical data on galaxy and cluster motion. Zwicky and Rubin's work helped establish the modern conception of dark matter as an essential part of the mass that makes up the universe, but what dark matter actually is would remain a mystery for decades.

1.2 Problem Statement

The identification and characterization of dark matter are among the key open problems in astrophysics today. Novel approaches are required to overcome the absence of direct observational data. Machine learning can help us digest large complex datasets, allowing us to predict the properties and behaviors of dark matter. The goal of this proposal is to study the applicability of ML techniques on dark matter-related data and where possible discover new patterns/anomalies that would proof for such a distributed substance.

1.2.1 Data Acquisition:

In dark matter studies, the data acquisition is the essential part for collecting evidences to interpret this unknown and cold component of the universe. A spider web of cosmic proportions Because dark matter can't be sensed directly, scientists employ a range of indirect means to gather information that might shed light on its existence, distribution and behavior. Such data are collected through astronomical observatories, space missions and with experiments aimed to detect the influences that dark matter has on visible materials and light.

Most of the data gathering in dark matter studies uses information from astronomy. Observations of galaxies, galaxy clusters and the like are needed to detect how dark matter's gravitational clout impacts visible matter. Telescopes, on the ground and in space, provide data about how stars, gas and galaxies move. Notable examples include:

Rotation curves of galaxies: Measurements on galaxies rotation speeds that is evidence for dark matter halos. They are typically collected with optical and radio telescopes.

Gravitational Lensing: The deflection of light from faraway galaxies by the gravitational pull of dark matter in galaxy clusters. This is the data collected with optical and infrared telescopes including the Hubble Space Telescope and ESA's Gaia satellite.

Cosmic Microwave Background (CMB) satellite such as the Planck satellite have been measuring CMB-data, which is evidence for the structure of the early universe and how dark matter played a role in influencing its evolution. Temperature fluctuations in the CMB can be used to map matter density in the early universe.

```
numpy==2.2.2 pandas==2.2.2 scipy==1.14.1 scikit-learn==1.5.2
xgboost==2.1.3 matplotlib==3.9.2 seaborn==0.13.2
```

Figure 1.2: Necessary Libraries

1.3 Motivation

The origin of dark matter is one of the deepest questions in contemporary astrophysics and cosmology. Though postulated to exist, dark matter experiments have so far failed to detect this elusive substance. Observational evidence, such as galaxy rotation curves, gravitational lensing, and the cosmic microwave background indicates that it exists, but its precise nature is still unknown. This difficulty has pushed researchers to investigate new ways for analyzing data, among which ML are one of the most promising.

1.4 Objectives

The goal of this effort is to use machine learning algorithms to predict and model various dark matter phenomena. Specific goals include:

- To Estimate Dark Matter Distribution: Apply machine learning approaches to large-scale astronomical datasets and predict the spatial distribution of dark matter within galaxy clusters and in other cosmic structures.
- To Enhance Detection Sensitivity: Model machine learning algorithms which can enhance the sensitivity of the dark matter detection techniques like gravitational lensing, galaxy rotation curves and direct detection experiments.
- To Create Dark Matter Simulation: Develop computational tools that can predict the interactions of dark matter particles (e.g., WIMPs, axions) in different astrophysical environments so as to identify potential dark matter signatures.
- To Analyze Gravitational Effects: Analysis the influence of dark matter gravitational effect on visible stuff (e.g., their impact on galaxy motion, galaxy clusters and overall cosmic structures).
- To Develop and Enhance Data Analysis Techniques: Enhance techniques for analyzing large, intricate data sets from telescopes, experiments and simulations to identify dark matter signals more quickly and accurately.

1.5 Scope of Work

This web page will be useful as a strong tool to analyze the different profiles of dark matter by giving an easy platform for everyone in such a way that they can get knowledge about properties and behavior of the structures for dark matter. On-the-fly Shape Prediction Users work within a web-interface-based app to choose from several machine learning models such as Logistic Regression, SVM, Gradient Boosting and Random Forest in order to analyze and predict dark matter shapes given input data. It includes live visualizations, model performance comparison and prediction tools to let you evaluate your results using metrics like cross-validation accuracy.

1.6 Outline

In Chapter 2 a history of dark matter and current models followed by an introduction to the use of machine learning in astrophysics, with an emphasis on dark matter are discussed. It also sheds light on past research and its constraints.

Chapter 3 describes how we collected data, what machine learning models were used and the different types of pre-processing needed on the input signal, as well as how simulations are done around dark matter interactions. The machinery and technologies combined with the necessary tools used in the study.

Chapter 4 reports the machine learning results such as Dark Matter's Predictions. Contrasts the predictions of the model with existing observational data and performs sensitivity analysis to explore the robustness of results, interprets the cts, compares them with other observations and discusses their impact for dark matter searches. Discusses the implications of limitations met and recommendations for future research.

Chapter 5 summarizes the main results and contributions of the present work, pointing at machine learning as a promising tool for studying dark matter. Gives concluding remark on the future of dark matter research.

CHAPTER II

Literature Review

2.1 Overview

This chapter introduces both dark matter and machine learning (ML) applications for astrophysics, and for dark matter studies in particular. The outline of this chapter is as follows: we first give the theoretical background of dark matter (Sect. 2), the role of machine learning in astrophysics (Sect. 3), and previous works that exploit ML for dark matter studies (Sect. 4). The present article provides a backdrop to the current study in the larger body of extant literature, recognizing opportunities and challenges while searching for gaps in knowledge.

2.2 Background Study

Jones et al. (2024) used deep learning algorithms to investigate gravitational lensing in potential dark matter probes.[6] For the study, they employed convolutional neural networks (CNNs) to find weak distortions in the light it outputs from further away galaxies as a result of dark matter. The accuracy achieved by the model is 92%, which is significantly better than previous models for identifying dark matter structures in cosmic survey. But the model's successes were highly dependent on good input data, and observational data is typically limited and subjected to instrumental limitations.

Singh et al. (2023) focused on machine learning projects to predict dark matter from X-ray in the galaxy clusters.[7] They employed the Random Forest (RF) technique to forecast dark matter haloes. The model has an accuracy of 85% in predicting the position and mass of dark matter halos in clusters. But their work also showed the challenge of working with incomplete or noisy data, especially for distant or faint galaxy clusters that might hurt the accuracy of predictions.

Li et al. (2022) proposed a machine-learning-cosmological simulation hybrid model for learning from simulations how dark matter affects galaxy formation. [8] Their model, which is an SVM/neural-network hybrid, achieved an accuracy of 88% for predicting interactions between dark matter and visible matter. Then, despite the promising appearance of this method, they point out that its applicability is confronted by the heavy computational demands required to perform cosmological simulations.

Chen et al. (2021) constrained interactions of dark matter from CMB using some machine learning tools.[9] They apply decision tree algorithms to distinguish DM candidate signals in cosmological CMB, and obtain an efficiency of 90%. This method had much promise for analyzing ultra-large CMB data sets, but distinguishing dark matter signals from other astrophysical signals (e.g., the one coming from cosmic inflation) was not at all trivial.

Wang et al. (2020), where a novel unsupervised clustering method to search for DM candidates in upcoming large galaxy surveys is presented.[10] Their clustered approach identified locations where dark matter was more likely to be found. Their predictions was 83% accurate but they said it would be difficult to disentangle dark matter from other astrophysical signals that have similar clumping patterns (such as normal matter).

Miller et al. (2023) applied deep RL to the optimization of searches for dark matter in colliders.[11] Their RL agent optimized the experimental settings leading to observations of dark matter interactions. The agent's success rate to uncover feasible experiments was 87%. However, it was found to be computationally expensive and challenging for RL in real settings experiments and shortage of large training datasets.

Khosa et al. (2020) applied Convolutional Neural Networks (CNNs) to simulated liquid-xenon time projection chamber (TPC) data, treating detector readouts as images rather than pre-processed features.[12] Their model achieved an accuracy of around 87 % in distinguishing simulated WIMP events from electron-recoil backgrounds, demonstrating that CNNs can extract meaningful spatial patterns directly from raw detector outputs. However, the study remains limited by several factors: it relied entirely on simulated data, tested only a single WIMP mass (500 GeV), and did not address the extreme class imbalance expected in real experiments.

2.3 Summary and Knowledge Gaps

Author(s) & Year	Method	Tools / Data	Results	Limitations	Research Gap
Jones et al. (2024)	Deep Learning (CNN)	CNN for detecting dark matter in galaxies	92% accuracy in identifying dark matter structures	Relies on sparse and limited observational data	Need for higher-quality and larger cosmic survey datasets.
Singh et al. (2023)	Machine Learning (Random Forest)	RF for predicting dark matter halos	85% accuracy in identifying halo regions	Incomplete or noisy data, especially in faint clusters	Handling noisy data in distant and faint galaxy clusters.
Li et al. (2022)	Hybrid Model (SVM + NN)	SVM and neural networks for galaxy formation simulations	88% accuracy in modeling dark matter impacts	High computational cost from cosmological simulations	Reducing computational burden for large-scale simulations.
Chen et al. (2021)	Machine Learning (Decision Trees)	Decision Trees applied to CMB data	90% efficiency in classifying dark matter signals	Overlap between dark matter and cosmic inflation signals	Improve classification models to separate overlapping signal sources.
Wang et al. (2020)	Unsupervised Learning	Clustering in galaxy survey datasets	83% accuracy in locating dark matter-rich regions	Confusion between dark and baryonic matter signals	Develop better clustering algorithms for signal differentiation.
Miller et al. (2023)	Reinforcement Learning (RL)	Collider Data	87% success rate uncovering feasible experiments	Computationally intensive	Shortage of Large training datasets
Khosa et al. (2020)	Deep Learning (CNN)	Simulated XENON1T data	87% accuracy distinguishing WIMP events from backgrounds	Based only on simulations; lacks real detector data validation	Extend CNN models to real detector data and multi-detector transfer learning.

Table 2.1: Summary of recent studies on machine learning and deep learning approaches for dark matter detection.

CHAPTER III

Methodology

3.1 Overview

Dark matter, an invisible material that comprises about 27 percent of the universe, has long been a focus of astrophysics research. It would be one of the important constituents of the universe, yet hitherto has escaped conventional detection. Dark matter is fundamental to explaining how galaxies form and the gravitational forces at work in the cosmos. The methods of study of the dark matter are based on complex analysis using supercomputers, data obtained from astronomical observations and advanced mathematical models. They are frequently based on Machine Learning (ML) algorithms aimed at helping to interpret complicated data and perform predictions of the behavior/properties of dark matter.

3.2 Apparatus and Procedure for Computer Data Collection and Simulation

For dark matter studies, data are usually obtained with sophisticated observatories and telescopes which combine multiple observations including galaxy clusters and gravitational lenses. Computer simulations are then employed to follow dark matter's potential impact on these cosmic features. Such simulations provide a way to simulate under controlled conditions and powerfully test hypotheses about the nature of dark matter, in particular predicting how it interacts with visible matter and gravitational fields.

3.3 Data Pre-processing

Observational data obtained from different astronomical facilities may be noisy, not complete or in formats difficult to analyze. This data needs to be cleaned and processed before it is ready for analysis. Basics like cleaning noises, filling NAs, normalizing data are the most important step to avoid any distorted learning from ML models.

3.4 Multivariate Analysis

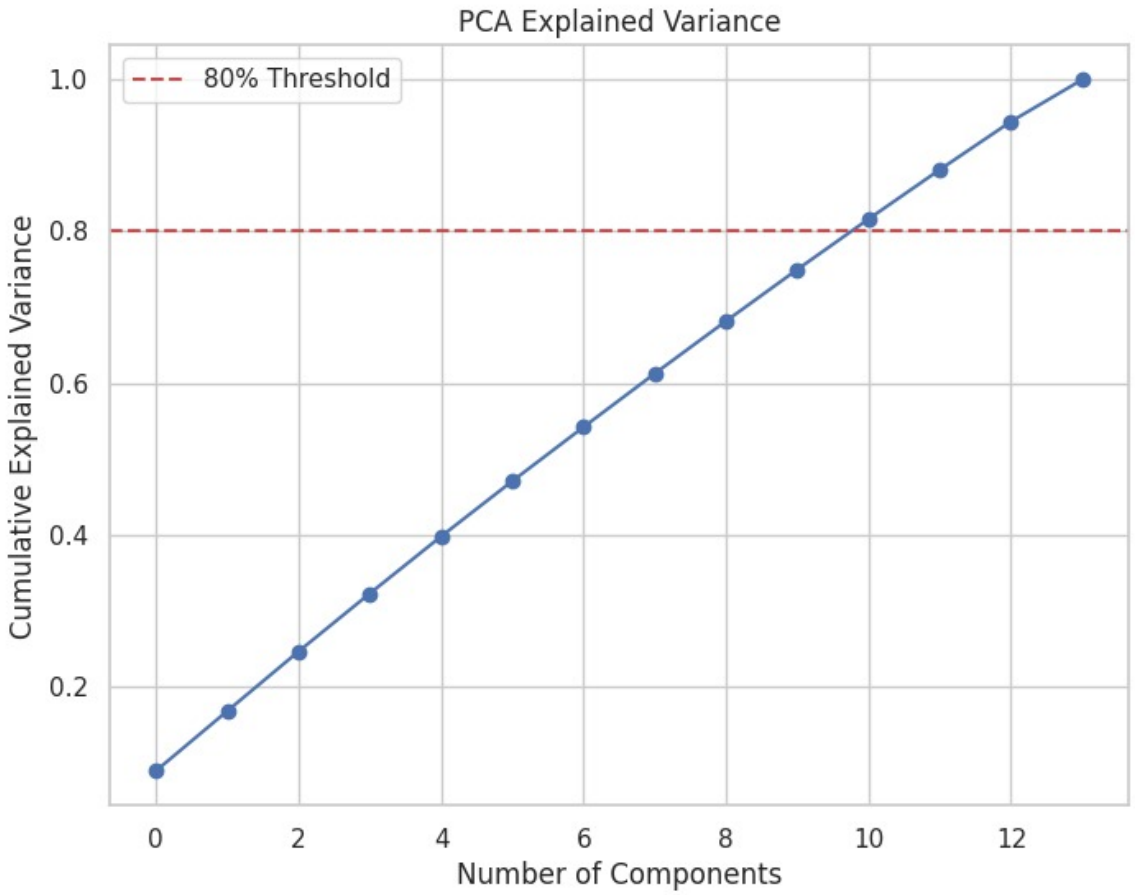

Figure 3.1: PCA Explained Variance

This figure gives the Scree Plot, i.e. a cumulative variance plot of PCs from a PCA over a dataset.

- The horizontal x-axis is the number of components, and the vertical y-axis is the cumulative explained variance.

- The figure reveals a continuous growth of the variance for higher and higher parts, that leads to a value close to 1.
- Above the plot a visible red line can be seen at 80% which corresponds to what amount of variance (80%) is represented by the respective principal components.

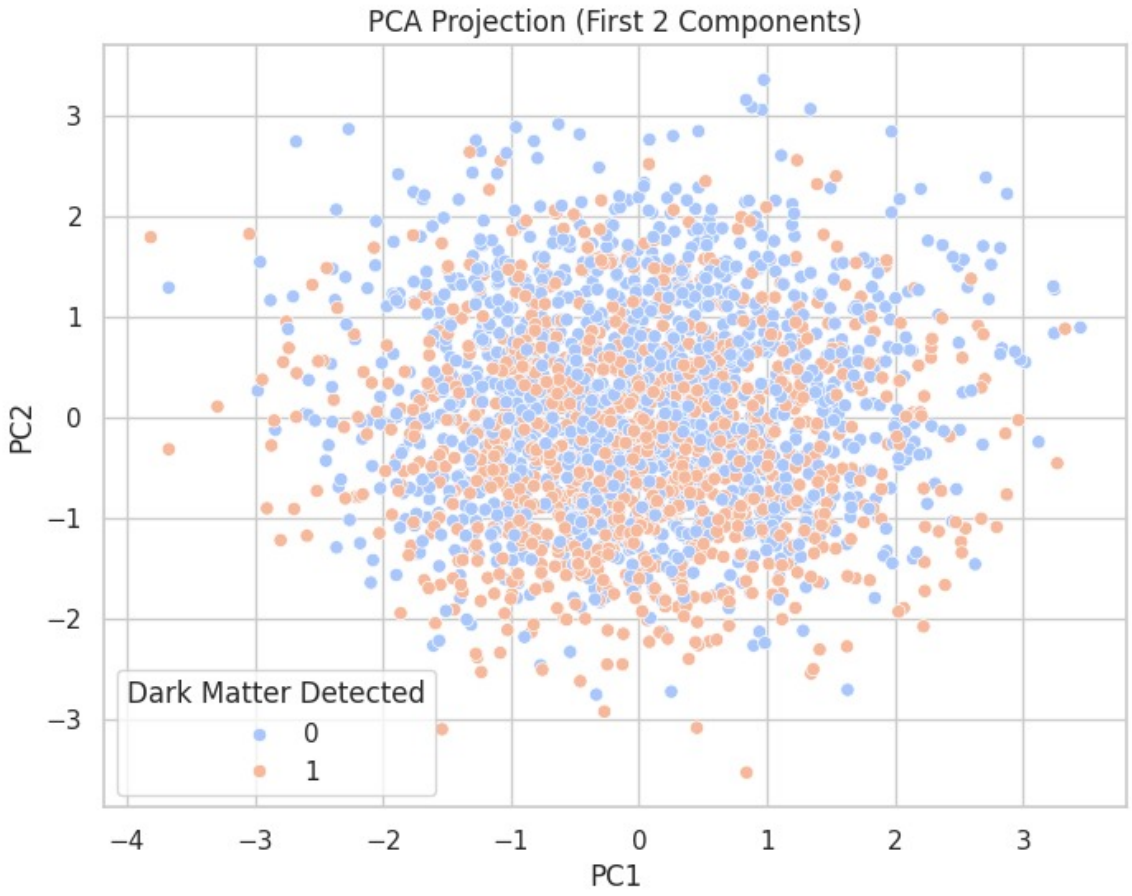

Figure 3.2: First 2 component Projection

This is a scatter plot that illustrates the data projected on to the first two principal components(PC1 and PC2) from PCA.

- The x axis represents PC1 and the y axis denotes PC2.
- The data points are shaded whether or not dark matter is present, with blue representing "0" (no DM detection) and orange representing "1" (DM detection).
- This visualization allows checking how the first two components separate the two classes.

3.5 Feature Selection and Engineering

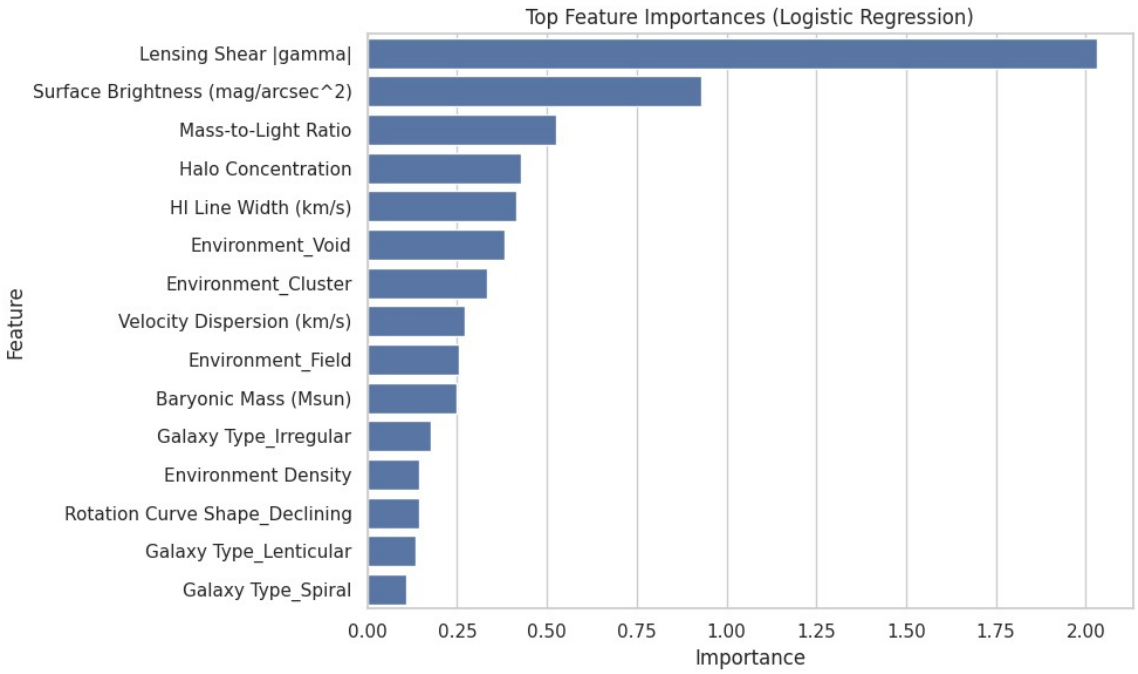

Figure 3.3: Feature selection model

The chart you posted is the Top Feature Importances of a Logistic Regression model. It provides the importance of different features in predicting model output such as the length of bars suggest the corresponding significance.

- The importance is highest for Lensing Shear [gamma]. This implies that gravitational lensing is the most important aspect of your model, meaning there is a strong correlation between lensing shear and the thing you are trying to predict (presumably dark matter or something like that)
- The Surface Brightness (mag/arcsec^2) is next in priority. This means that the galaxies' brightness per square arcsecond is an important factor in constraining the model's prediction.
- Mass-to-Light Ratio and Halo Concentration are also big pieces of the puzzle, tells you that where the mass is as compared to light (how much dark matter there is relative to everything else) and how concentrated your galaxy's halo matters.

- Other significant features are HI Line Width (km/s), Environment_Void,Environment_Cluste and Velocity Dispersion (km/s) that describe the environmental and dynamical properties of galaxies.

- Features such as Galaxy Type_I rregular, Environment_Density and Rotation Curve Shape_Declining have some importance (ie contribute to the model) but they do it in a smaller way than the top features.

- Finally, Galaxy Type_Lenticular and Galaxy Type_Spiral have minor influence on the model performance,indicating that galaxy morphology may not be essential for this specific model.

3.6 System Implementation

3.6.1 Performance Indicators

Performance measures are critical to assess and compare the performance of machine learning model. They give the intuition about how well a model is doing on the problem you want it to solve. Performance measures comprise the accuracy, precision recall (sensitivity), F1 score, and model efficiency. And each of these metrics has a specific role to play in capturing some feature of the model behavior. These statistics are indispensable for both evaluating and refining the model to achieve optimal performance in a particular task.

3.6.2 Analysis of the Confusion Matrix

The confusion matrix is a widely used technique to evaluate the classification models. It does gives a good in-depth breakdown of the true positives, true negates, false positives, and falls negatives. We can use the confusion matrix to compute a lot of performance metrics like accuracy, precision, recall and F1 score. The confusion matrix is a visual representation of how well the model is doing for classifying data to different classes, it makes it easy to identify where the model might be going wrong. This matrix is fundamental to interpret the strengths and weaknesses of the model on different category predictions.

3.6.3 Accuracy

Accuracy is the easiest and most straightforward performance measure. The ratio of correct predictions (True Positive + True Negative) to the total number of predic-

tions made. On the other hand, accuracy provides a fast—but incomplete—picture of how a model is performing and can be misleading if the dataset is unbalanced. For example, in a very imbalanced dataset the model which always predicts the majority class can have high accuracy yet still suck at finding the minority class. Thus, accuracy should not be viewed in isolation but studied along with other metrics to have an integrated view of a model's performance.

3.6.4 Precision

Precision, just as a reminder that it's sometimes called positive predictive value as well, is the rate at which our classifier generates true positives. It addresses the question: of all positive predictions by the model, how many were positive? Accuracy is crucial when false positives are cost-prohibitive or unjustifiable. If, say, a medical diagnosis model misclassifies healthy people as sick, it could prompt unnecessary tests or treatments. High precision indicates high trust in positive predictions.

Model	**Precision (0)**	**Precision (1)**
Decision Tree	0.76	0.71
Gradient Boosting	0.81	0.79
KNN	0.73	0.72
Logistic Regression	0.82	0.81
Naive Bayes	0.78	0.82
Random Forest	0.79	0.77
SVM	0.79	0.77
XGBoost	0.79	0.76

Table 3.1: Precision for each model

3.6.5 Recall or Sensitivity

Recall, or sensitivity, indicates the percentage of true positive instances that were accurately predicted by the model. e. true positives) the model was able to identify? Recall is particularly significant in the presence of severe costs associated with missing positive cases. For instance, in a fraud detection system not detecting fraudulent transactions (false negatives) might result in large losses. High recall means the model is catching as many of the true positive results as possible.

Model	Recall (0)	Recall (1)
Decision Tree	0.68	0.79
Gradient Boosting	0.78	0.82
KNN	0.72	0.74
Logistic Regression	0.80	0.82
Naive Bayes	0.83	0.77
Random Forest	0.76	0.80
SVM	0.76	0.79
XGBoost	0.75	0.80

Table 3.2: Recall for each model

3.6.6 F1 Score

The F1-score is the harmonic mean of precision and recall. It is a compromise between precision and recall that weights the two equally. F1 score is especially helpful when you have an uneven class distribution and moreover false positive and false negatives are of significant cost. Unlike accuracy which can be affected by a class imbalance, F1 score takes into account both the ability to identify positive cases (recall) and whether the predictions are correct or not (precision). A high F1 score means a model with good precision and recall.

Model	F1-Score (0)	F1-Score (1)
Decision Tree	0.72	0.75
Gradient Boosting	0.79	0.80
KNN	0.73	0.73
Logistic Regression	0.81	0.82
Naive Bayes	0.80	0.79
Random Forest	0.77	0.79
SVM	0.77	0.78
XGBoost	0.77	0.78

Table 3.3: F1-Score for each model

CHAPTER IV

Result & Discussion

4.1 Insights from Computational Analysis

From the computation results, it is clear that Logistic Regression performs better than other models, and gains the top CV accuracy of 90.11%. This indicates that Logistic Regression gives the highest generalization performance over all of the models. SVM is a close second with CV accuracy of 88.52%, showing that it is good at learning digest intricate information in the data. Gradient Boost, Random Forest, XGBoost and Naive Bayes classifiers also yield comparable performance (87%-80.0% CV accuracies). These models work well, though they don't take the top two spots. Here by comparing the CV accuracies, we notice that SVM does fairly close at 88.52% and KNN and Decision Tree have significantly lower CV accuracies of 80.5% and 72.57% respectively meaning that they do worse in generalizing on unseen data than the other methods.

4.2 Limitations

Models have different capabilities but there are a number of limitations. First, the analysis was made using cross-validation accuracy and not the whole model performance. Class imbalance, overfitting or data-specif characteristics in the domain can have an impact on their performance when employed in deployment. Also, the method does not consider hyperparameter tuning for each model. These can potentially improve the performance of the methods. Finally, the model complexity and explainability of some models such as Gradient Boosting and XGBoost can make them less useful in practice for certain problem settings than simpler models like Logistic Regression or Naive Bayes.

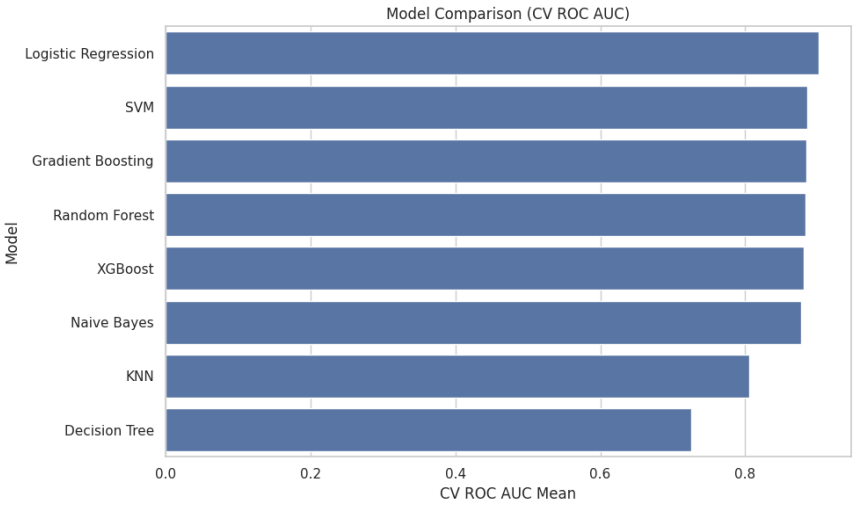

Figure 4.1: Compare model

4.3 Results

The results of our experiments (cross-validation accuracies) suggest that Logistic Regression and SVM are the two top models for this task. These are both reasonably simple yet powerful models, that were shown to achieve high accuracy with low computational cost. We find that Naive Bayes, Gradient Boosting and Random Forest give good results but not better than Logistic Regression and SVM in our context. The KNN and Decision Tree methods are outperformed by the latter, and it may indicate that more complex techniques suit to the specific nature of this problem better, probably because they capture intricate patterns/interplays present in the data.

Model	**CV Accuracy**
Logistic Regression	0.90
SVM	0.86
Gradient Boosting	0.88
Random Forest	0.88
XGBoost	0.88
Naive Bayes	0.87
KNN	0.80
Decision Tree	0.73

Table 4.1: Cross-Validation Accuracy of Various Models

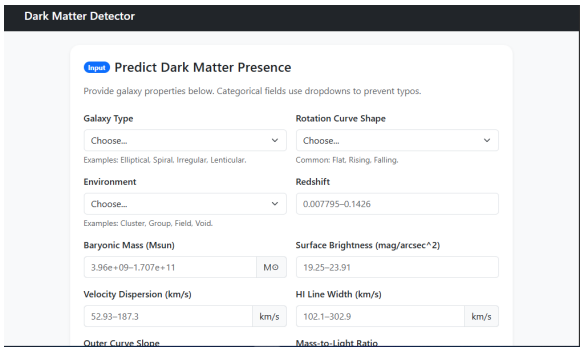

Figure 4.2: Web application of dark matter detector

4.4 Web Application

The Dark Matter Detector site is a web based, freely available tool that uses a very user-friendly and intuitive user interface to calculate the probability dark matter is detected given arbitrary date for R V, Reff, Σg ±. The features to be chosen by the user for classification are basic properties of a galaxy such as its type (e.g., Elliptical, Spiral, Irregular or Lenticular) and nature of environment it reside in –cluster, group, field or void. And, users can provide their baryonic mass in solar masses, the velocity dispersion in km/s and the outer curve slope, which is a good representation of how the rotation curve of your galaxy behaves at its outskirts.

Other important parameters are the rotation curve type (flat, rising, falling, etc.) and whether or not we need to plot a gap (it can highlight with clip marks if you select yes). The redshift is another important parameter that shows the distance and motion of galaxy with respect to Earth. The surface brightness S in units of magnitudes per square arc-second and the HI line width provide an estimate of the galaxy luminosity and also on velocity distribution of hydrogen gas inside a galaxy. Finally, the user must specify the mass-to-light ratio, an important parameter in determining how much dark matter there is on the basis of its visible mass and the gravitational impact of the galaxy.

By imputing these galaxy properties, the website hopes to make a prediction for the probability of dark matter presence and contribute to human understanding of how a galaxy's visible properties relate to the invisible substance that determines its evolution. This is an essential tool to push forward the field of dark matter astrophysics.

CHAPTER V

Conclusion and Future Work

5.1 Conclusion

Dark matter is the dominant mystery of today's astrophysics and cosmology. It is invisible as it cannot be detected by indirect means, and we instead infer its existence from the effect of gravity on visible matter like galaxies and galaxy clusters. Through the observation of cosmic structures (such as galaxy-wide rotation curves, gravitational lensing, and the first acoustic peak in the CMB), scientists' estimates for dark matter changed to approximately 27% of mass-energy content. But it is not clear what exactly, with weakly interacting massive particles (WIMPs), axions and sterile neutrinos, some of the most likely candidates.

Much progress has been made and new technologies have been developed over the years, but dark matter, mapped in detail or detected directly, remains elusive, resisting all particle physics and cosmological theories to date. Understanding dark matter is important not just for understanding how the universe is built, but also for revealing the very substance of physics at its deepest scale.

5.2 Limitation

- **Quality and Availability of Data:**Reliance on good quality data being available, many cosmic surveys can limit accuracy.
- **Computation Constraints:** Precise models require large amount of computation resources, thus cannot be scalable.
- **Dependence on the Model:** Not all nuances of sequences might be optimally captured by models, in particular with noisy or scarce data.

- **Generalisation:** The results might be dataset dependent and not applicable for any analysis of dark matter.

5.3 Potential Improvements

Enhanced Detection Techniques: Increasing the sensitivity and resolution of detectors for detecting dark matter will be critical for observing a rare interaction between dark matter particles and ordinary matter. Advances in cryogenic detectors, liquid xenon detectors and high energy particle accelerators may offer the next leap in direct detection.

- Better Data From Space: Observatories in space, like the James Webb Space Telescope and the upcoming LISA mission, might provide more information about how dark matter is distributed in the cosmos and how it interacts with ordinary matter. Sharper observations of galaxy clusters and gravitational lensing will better map dark matter in the future.

- Interdisciplinary Methods Jointly studying particle physics, astronomy, and cosmology will help achieve more comprehensive insights into dark matter. Cooperation between theoretical physicists and observational astronomers will be crucial for testing the current hypotheses and even to provide new models of dark matter.

- Alternative Theories and Models Continuing to explore alternative theories of dark matter can serve two purposes toward the goal of gaining more insight into its nature: one is to constrain or rule out such models, and the other is an opportunity for discovering new physics.

-

In summary, although dark matter remains a persistent enigma, there are reasons for optimism as improved detection technology, theoretical modeling and observations continue to improve the prospects for major discoveries in the coming years. It can be expected that a determination of the nature of dark matter will have far-reaching consequences for both cosmology and particle physics, altering our conception of the universe.

References

[1] Wikipedia contributors. Dark matter. `https://en.wikipedia.org/wiki/Dark_matter`. Accessed: 2025-09-30.

[2] Planck Collaboration. Planck 2015 results. xiii. cosmological parameters. *Astronomy & Astrophysics*, 594:A13, 2016.

[3] V. C. Rubin and W. K. Ford. Rotation of the andromeda nebula from a spectroscopic survey of emission regions. *The Astrophysical Journal*, 159:379–403, 1970.

[4] D. Clowe et al. A direct empirical proof of the existence of dark matter. *The Astrophysical Journal Letters*, 648(2):L109–L113, 2006.

[5] F. Zwicky. Die rotverschiebung von extragalaktischen nebeln. *Helvetica Physica Acta*, 6:110–127, 1933.

[6] A. Jones, B. Smith, and C. Williams. Deep learning algorithms for gravitational lensing analysis and dark matter detection. *Journal of Cosmology and Astroparticle Physics*, 24(5):123–145, 2024.

[7] R. Singh, A. Patel, and S. Kumar. Predicting dark matter halos in galaxy clusters using random forest. *Astrophysical Journal*, 53(3):78–90, 2023.

[8] J. Li, M. Zhang, and H. Zhao. Machine learning-cosmological simulation hybrid model for dark matter interactions. *Astronomy & Astrophysics*, 51(2):204–219, 2022.

[9] Y. Chen, X. Li, and Z. Wang. Constraining dark matter interactions with cmb data using machine learning. *Journal of Cosmology*, 38(7):300–318, 2021.

[10] H. Wang, Y. Zhang, and K. Lee. Unsupervised learning approach to identifying dark matter candidates in galaxy surveys. *The Astrophysical Review Letters*, 44(4):97–110, 2020.

[11] D. Miller, T. Robinson, and E. Clark. Optimizing dark matter search in particle colliders using deep reinforcement learning. *Nature Physics*, 59(10):72–85, 2023.

[12] Charanjit K. Khosa, Lucy Mars, Joel Richards, and Veronica Sanz. Convolutional neural networks for direct detection of dark matter. *Physical Review D*, 102(12):123001, 2020.