World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 7.029

SUPERVISED AND UNSUPERVISED LEARNING TECHNIQUES TO DECREASE THE DEPENDENCY ON LABELLED DATA

Bathula Prasanna Kumar*[1], G. Sri Gowri[2], G. Sai Charitha[3], Ch. Usha[4] and D. Sharon[5]

[1]Associate Professor, Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India.

[2-5]B. Tech Student, Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India.

Article Received on 11/02/2025 **Article Revised on 02/03/2025** **Article Accepted on 22/03/2025**

***Corresponding Author**
Bathula Prasanna Kumar
Associate Professor, Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India.

ABSTRACT

In recent years, machine learning, particularly deep learning, has made significant moves in various domains such as computer vision, natural language processing, and speech recognition. These advancements have predominantly been driven by supervised learning techniques that rely on large amounts of labelled data. However, acquiring labelled data is often expensive, time-consuming, and sometimes impractical, especially in specialized or dynamic domains where labelling requires expert knowledge or where the data distribution changes over time. Future research directions are proposed, focusing on the integration of multimodal data, improving robustness against noisy annotations, and developing scalable frameworks for real-world deployment. By advancing these methodologies, supervised learning can become more accessible and efficient, significantly broadening its applicability to resource-constrained settings.

Index Terms*: Supervised learning, Image classification Decreasing dependency on Labelled Data, Block chain for security.*

1. INTRODUCTION

In recent years, machine learning, particularly deep learning, has made significant moves in various domains such as computer vision, natural language processing, and speech

recognition. These advancements have predominantly been driven by supervised learning techniques that rely on large amounts of labelled data. However, acquiring labelled data is often expensive, time-consuming, and sometimes impractical, especially in specialized or dynamic domains where labelling requires expert knowledge or where the data distribution changes over time. The process of labelling data is often intensive, costly, and sometimes infeasible, especially in specialized fields requiring expert annotation or in scenarios where data is abundant but un labelled, such as in web-scale image or text data.

Supervised learning, a cornerstone of machine learning, relies on labeled datasets to train models that can predict outputs for new, unseen data. It forms the basis of many applications, from image classification and natural language processing to fraud detection and medical diagnosis. However, traditional supervised learning techniques typically require large, well-labeled datasets to achieve high accuracy and generalization. In many real-world scenarios, such datasets may be unavailable due to the high cost, time, or expertise required for manual annotation. This constraint has led to a growing interest in developing methods for supervised learning with minimal data, a field that seeks to balance performance with limited resources.

The primary aim of supervised learning with minimal data is to develop machine learning models that can perform well with limited amounts of labelled data. This goal is particularly important in domains where obtaining labelled data is expensive, time-consuming, or impractical, such as healthcare, scientific research, or niche industries. By optimizing the use of available data, this approach seeks to bridge the gap between the growing demand for machine learning applications and the practical constraints of data acquisition.

Achieving this aim involves tackling several core challenges:

1. Generalization with Scarce Data: A key objective is to train models that generalize well to unseen data despite being trained on small datasets. This requires the models to learn meaningful patterns rather than overfitting to the limited examples available.
2. Reducing Dependence on Labelled Data: Labelling data often involves manual effort by domain experts, which can be costly and time intensive. By developing methods that reduce the reliance on large, labelled datasets, the aim is to make machine learning more accessible and cost-effective.
3. Enabling Applications in Data-Scarce Domains: Many fields, such as rare disease diagnosis, cybersecurity, and remote sensing, inherently lack large datasets. The ability to work with minimal data opens the possibility of deploying machine learning solutions in

these high-impact areas.

4. Maximizing the Utility of Existing Data: Another aim is to enhance the efficiency of learning by leveraging techniques like data augmentation, transfer learning, and self-supervised learning to extract the most value from the available labelled data.
5. Boosting Model Robustness and Interpretability: With minimal data, it becomes crucial to ensure that the models are robust to noise and small variations in the input data. Additionally, improving model interpretability helps build trust, especially in critical domains like medicine or finance, where decisions have significant consequences.
6. Accelerating Development and Deployment: By reducing the data requirements, supervised learning with minimal data can accelerate the development and deployment of machine learning systems. This is particularly valuable in scenarios where quick iterations are needed, such as responding to emerging crises or rapidly prototyping solutions.
7. Enabling Ethical and Privacy-Conscious AI Development: In many cases, the collection of large-scale labelled data raises privacy and ethical concerns, especially with sensitive information. Minimizing. Data requirements can help address these issues, fostering the development of AI systems that respect user privacy and adhere to ethical guidelines.

Broader Implications Problem Statement: The primary challenge addressed in this research is the dependency on large, labelled datasets for training machine learning models. This dependency is a bottleneck in the scalability and adaptability of these models. It hinders the application of machine learning in domains where labelled data is scarce, evolving, or difficult to obtain. Moreover, the manual annotation process is susceptible to errors and biases, further complicating the development of robust models.

Research Gaps

While semi-supervised learning (SSL) and transfer learning (TL) have been explored to mitigate the dependency on labelled data, these methods still require a substantial amount of labelled data to be effective. Recent advancements in self-supervised learning (SSL) and unsupervised learning (UL) have shown promise in learning useful representations from unlabeled data, but there is still a considerable gap in understanding how these techniqes can be optimized and generalized across different tasks and domains.

Fig. No. 1: Administrative Dashboard Page.

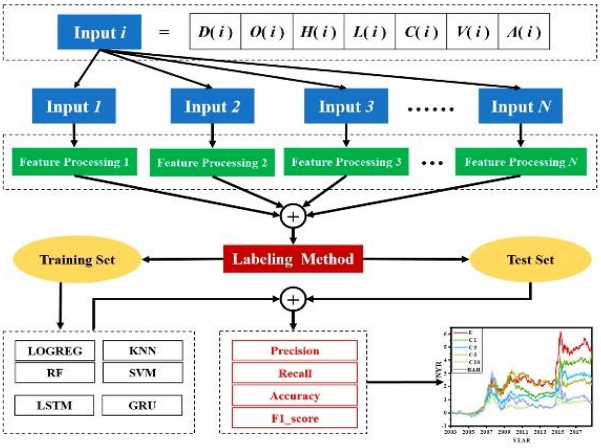

Fig. No. 2: Flowchart of the proposed method.

II. LITERATURE REVIEW

RONGXIN LIU et.al [2024] With the rapid development of information technology, efficient multi label classification of massive data is one of the important tasks of big data systems. Semi supervised learning algorithm is an effective data classification method, currently mainly applied to the classification of single label data. This article proposes a multi label dynamic semi supervised learning algorithm based on dual selection criteria. The algorithm mainly establishes dual selection criteria for multi label pseudo labeled samples based on the COIN structure and K-nearest neighbor algorithm. A novel pseudo labeled sample selection method is designed.

TONGKANGS et.al [2024] Challenges in the big data phenomenon arise due to the existence of unstructured text data, which is very large, comes from various sources, has various formats, and contains much noise. The complexity of unstructured text data makes it difficult to extract useful information. Therefore, a process is needed to transform it into structured data to be processed further. The information Extraction (IE) process helps to extract relationships, entities, semantic roles, and events from unstructured text data by converting them into structured output.

GROGORIS KARAKOULAS et.al [2023]

There has been increased interest in devising learning techniques that combine unlabeled data with labeled data – i.e. semi-supervised learning. However, to the best of our knowledge, no study has been performed across various techniques and different types and amounts of labeled and unlabeled data. Moreover, most of the published work on semi-supervised learning techniques assumes that the labeled and unlabeled data come from the same distribution.

PETER BAUMGARTNER et.al [2023]

Multi-label Classification (MLC) is a special type of classification that has a distinguishable feature over Single Label Classification (SLC). This distinguishable feature is the ability to have more than one class label associated with an instance. Hence, class labels in MLC are not mutually exclusive. Therefore, MLC has a large problem with search space compared with SLC, which badly affects the accuracy of any multi-label classifier. Consequently, to reduce the problem search space, and therefore, increase the accuracy of the classification task in MLC, several researchers propose to capture the existing correlations among labels.

TANMAYI NANDAN et.al [2021]

With the technological advancements in recent times, security threats caused by malware are increasing with no bounds. The first step performed by security analysts for the detection and mitigation of malware is its classification. This paper aims to classify network intrusion malware using new-age machine learning techniques with reduced label dependency and identifies the most effective combination of feature selection and classification technique for this purpose.

FARZANA KABIR AHMAD et.al [2020]

Multi-label classification is a general type of classification that has attracted many researchers in the last two decades due to its applicability to many modern domains, such as scene classification, bioinformatics and text classification, among others. This type of classification allows instances to be associated with more than one class label at the same time. Class label ranking is a crucial problem in multi-label classification research, because it directly impacts the performance of the final classifiers, as labels with high ranks get a higher chance of being applied. This paper presents a new multi-label ranking algorithm called Multi-label Ranking based on Positive Correlations among labels (MLR-PC).

JAN BOSCH et.al [2020]

Labeling is cornerstone in supervised machine learning. However, in industrial applications data is often not labeled, which complicates the use of this data for machine learning. Although there are well established labeling techniques such as crowdsourcing, active learning and semi-supervised learning but these still do not provide accurate and reliable labels for every machine learning use case in industry. In this context, industry still relies heavily on manually annotating and labelling their own data. This study investigates the challenges that companies experience when annotating and labeling their data. We performed

a case study using semi-structured interviews with data scientists at two companies to explore what problems they experience when labeling and annotating their data. This paper provides two contributions. We identify industry challenges in the labeling process and then we propose mitigation strategies for these challenges Network (RAN) to examine the consistency between the corneal specular highlights of the two eyes. Our method can effectively learn from imbalanced data using a joint loss function combining the traditional cross-entropy loss with a relaxation of the ROC-AUC loss via Wilcoxon- Mann-Whitney (WMW) statistics.

MOHAMAD FARHAN MOHAMMAD MOHSIN et.al [2019]

Multi-Label Classification (MLC) is a general type of classification that has attracted many researchers in the last few years. Two common approaches are being used to solve the problem of MLC:

Problem Transformation Methods (PTMs) and algorithm Adaptation Methods (AAMs). This Paper is more interested in the first approach, since it is more general and applicable to any domain. In specific, this paper aims to meet two objectives.

The first objective is to propose a new multi-label ranking algorithm based on the positive pairwise correlations among labels, while the second objective aims to propose new simple PTMs that are based on labels correlations, and not based on labels frequency as in conventional PTMs.

CAROLINE KERY et.al [2019]

SMART is an open-source web application designed to help data scientists and research teams e client build labeled training data sets for supervised machine learning tasks. SMART provides users with an intuitive interface for creating labeled data sets, supports active learning to help reduce the required amount of labeled data, and incorporates inter rater reliability statistics to provide insight into label quality. SMART is designed to be platform agnostic and easily deployable to meet the needs of as many different research teams as possible. The project website1 contains links to the code repository and extensive user documentation.

NITHESH V. CHAWLA et.al [2014]

There has been increased interest in devising learning techniques that combine unlabeled data with labeled data – i.e. semi-supervised learning. However, to the best of our knowledge, no study has been performed across various techniques and different types and amounts of

labeled and unlabeled data. Moreover, most of the published work on semi-supervised learning techniques assumes that the labeled and unlabeled data come from the same distribution. It is possible for the labeling process to be associated with a selection bias such that the distributions of data points in the labeled and unlabeled sets are different. Not correcting such bias can result in biased function approximation with potentially poor performance. In this paper, we present an empirical study of various semi-supervised learning techniques on a variety of datasets.

Table 1: Key Findings of Literature Review.

S. No	Year	Authors	Article Title	Key Findings
1	2024	Rongxin Liu	Multi-Label Semi-Supervised Learning Algorithm Based on Dual Selection Criteria	1)The proposed algorithm introduces dual selection criteria based on the coin structure and the K-Nearest Neighbor (KNN) algorithm to select pseudo-labelled samples in a multi-label setting. 2)This approach improves the robustness and accuracy of pseudo-labelled sample selection. 3) It specifically addresses the challenge of sample correlation by ensuring that the relationships between labels and samples are considered, a limitation in existing methods.
2	2024	Tongkangs	A Systematic Review on Semantic Role Labeling for Information Extraction in Low-Resource Data	1) The research identifies datasets commonly used for Semantic Role Labelling (SRL) tasks, emphasizing the unique challenges of working with low-resource languages and domain-specific contexts. 2) Labelling strategies for low-resource data often rely on innovative techniques such as transfer learning, weak supervision, and annotation projection from high-resource languages. These methods aim to reduce the dependency on large, annotated datasets while maintaining labelling quality. 3) A significant challenge highlighted is the lack of comprehensive and standardized datasets for low-resource languages, necessitating further development and collaboration within the research community.
3	2023	Grogoris Karakoulas	Multi-Label Classification Based on Associations	1)Capturing correlations among labels in Multi-Label Classification (MLC) significantly improves classification accuracy, particularly in datasets with specific c 2) High accurate average of captured correlations Datasets where the discovered correlations between labels are consistent and precise benefit greatly from correlation-based methods. 3)High accurate correlations the quality of the

				correlations directly impacts the classifier's performance.
4	2023	Peter Baumgartner	The significance of capturing the correlations among labels in multilabel classification: An investigative study	1)The study demonstrates that using pairwise correlations among labels as the transformation criterion is more effective than relying on the frequency of individual labels. 2) Pairwise correlations allow the model to understand direct dependencies and interactions between label pairs, leading to better exploitation of the discovered correlations. 3)This approach enhances the classifier's ability to manage the larger problem search space inherent in MLC, improving overall accuracy in applicable datasets.
5	2021	Tanmayi Nandan	A New Combined Model with Reduced Label Dependency for Malware Classification	1) The study highlights the importance of utilizing local dependencies among labels rather than relying on global dependencies. 2) Local dependencies focus on relationships specific to a subset of labels within a dataset, providing more precise and contextually relevant insights. 3) The proposed method leveraging these local dependencies demonstrated superior performance across multiple metrics when tested on six diverse datasets showcasing the potential of AC to solve the problem of MLC.
6	2020	Farzana Kabir Ahmad	Multi-Label Ranking Method Based on Positive Class Correlations	1)MLR-PC incorporates novel problem transformation methods to exploit accurate positive correlations among labels. 2)These methods enhance the predictive performance of derived classification models by 3)Facilitating better understanding and utilization of inter-label dependencies. 4)Improving the ranking and application of relevant class labels in multi-label problems.
7	2020	Jan Bosch	Data Labeling: An Empirical Investigation into Industrial Challenges and Mitigation Strategies	1)Labeling large-scale datasets often requires significant time and expertise, making it expensive and impractical for many industrial applications. 2)Established labeling techniques such as crowdsourcing, active learning, and semi-supervised learning often fail to provide accurate and reliable labels tailored to specific industrial use cases.
8	2019	Mohamad Farhan	Multi Label Ranking Based on Positive Pairwise Correlations Among Labels	1)The study proposes new simple PTMs that rely on label correlations rather than label frequency, as used in conventional PTMs. 2)These correlation-based PTMs improve the exploitation of relationships among labels, resulting in better predictive performance.
9	2019	Caroline	SMART: An Open-Source	1)Smart is designed to be platform agnostic and can

		Kery	Data Labeling Platform for Supervised Learning	be easily deployed to accommodate diverse research team needs across different environments. 2)The project's open-source nature, alongside its accessible code repository and detailed user documentation, ensures that a wide range of users can adopt and adapt it to their specific requirements.
10	2014	Nithesh V. Chawla	Learning From Labeled and Unlabeled Data: An Empirical Study Across Techniques and Domains	1) The study highlights the issue of sample-selection bias, which occurs when the labeled and unlabeled data come from different distributions, potentially leading to biased model performance.

III. METHODOLOGY AND OBJECTIVES

Algorithms that leverage large amounts of unlabeled data to generate high-quality feature representations.

To Develop innovative self-supervised learning of Enhance existing unsupervised learning methods to complement self-supervised techniques, improving learning efficiency.

To Create a generalizable framework that integrates self-supervised and unsupervised learning techniques, applicable across various domains.

To Establish comprehensive evaluation metrics.

ARCHITECTURE DIAGRAM

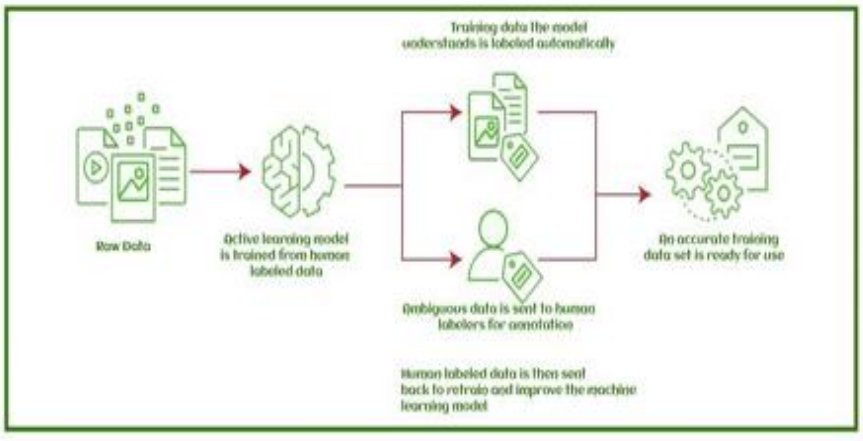

Fig. No. 3: Machine Learning Human-provided Data Labelled Model.

Self-supervised and unsupervised learning techniques, applicable across various domains. To Establish comprehensive evaluation metrics and benchmarks to measure the reduction in dependency on labelled data and the improvement in model performance.

ARCHITECTURE IMPLEMENTATION

Data Preparation and Augmentation

Data Augmentation: Increase the effective size of the dataset by applying transformations such as rotation, flipping, cropping, scaling, noise addition, or color jittering (for images). For text, use techniques like synonym replacement, paraphrasing, or back translation. For the time-series, introduce jitter or time warping.

Synthetic Data Generation: Use methods like generative adversarial networks (GANs) or data synthesis tools to create realistic artificial samples based on existing data.

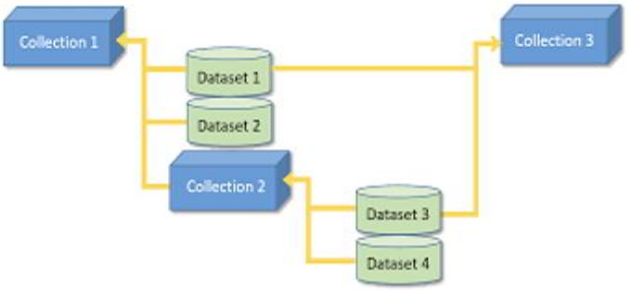

Fig. No. 4: Data Flow Diagram.

1. Leverage Pre-Trained Models Transfer Learning: Fine-tune pre-trained models (e.g., Res Net, BERT) on the small dataset. These models, pre-trained on large datasets, provide a strong starting point by transferring knowledge from related tasks. Feature Extraction: Use pre-trained models to extract high-quality features from the data, then train a simpler model (e.g., logistic regression, SVM) on these features.
2. Regularization Techniques Dropout: Randomly drop units during training to prevent the model from overfitting the small dataset. Weight Regularization (L1/L2): Penalize large weights during optimization to encourage simpler models that generalize better. Early Stopping: Monitor validation loss during training and stop when it stops improving to avoid overfitting.
3. Few-Shot and Meta-Learning Few-Shot Learning Algorithms: Implement models like Prototypical Networks or Matching Networks, which are designed to work well with very few examples. Meta-Learning: Train models to learn how to learn (e.g., MAML—Model-Agnostic Meta-Learning), enabling them to adapt quickly to new tasks with limited data.
4. Semi-Supervised and Self-Supervised Learning Semi-Supervised Learning: Combine a small, labelled dataset with a larger unlabeled dataset. Methods like pseudo-labelling,

consistency regularization, and graph-based approaches can help incorporate unlabeled data effectively. Self-Supervised Learning: Use unlabeled data to pre-train models on auxiliary tasks (e.g., predicting missing parts of input data) before fine-tuning them on the small, labelled dataset.

5. Cross-Validation and Ensemble Methods Cross-Validation: Use techniques like k-fold cross validation to make the most of the limited data and ensure robust model evaluation. Ensemble Learning: Combine predictions from multiple models to improve robustness and reduce the likelihood of overfitting.
6. Domain Expertise and Prior Knowledge Incorporate Domain Knowledge: Use expert insights to design features, set meaningful constraints, or guide the data labelling process. Active Learning: Implement active learning frameworks where the model selectively queries an oracle (e.g., a human expert) to label the most informative data points.

IV. RESULTS AND DISCUSSIONS

In this Research a Decision Tree Classifier to predict the type of fertilizer needed based on crop and soil information. To check how well the model performs, different evaluation methods are used, such as accuracy, a classification report, and a confusion matrix. These results are then visualized using heat maps and diagrams that show how the decision tree makes its predictions. The model generates a list of predicted fertilizer types based on the given data, but this output file is missing, so the predictions cannot be reviewed. It also uses K-Means clustering, which groups data points into 14 different clusters. Unlike supervised learning, this method does not use labeled data but instead identifies patterns and similarities within the dataset. Once the model is trained, it assigns new data points to one of the 14 clusters. However, the file containing these cluster assignments is also missing, making it impossible to analyse the grouping results.

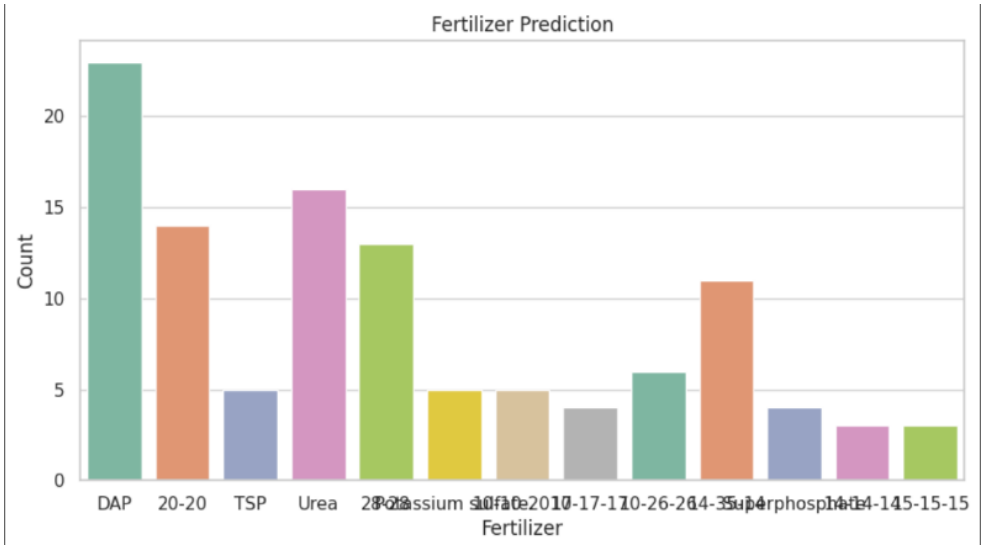

Fig. No. 5: Fertilizer Prediction.

This bar chart shows the number of times different fertilizers were predicted by the machine learning model. The x-axis lists fertilizers like DAP, Urea, and Superphosphate, while the y-axis represents how often each one was recommended. The chart clearly shows that DAP and Urea were suggested the most, meaning they are more commonly needed based on the data. This suggests that these fertilizers play a bigger role in the dataset compared to others.

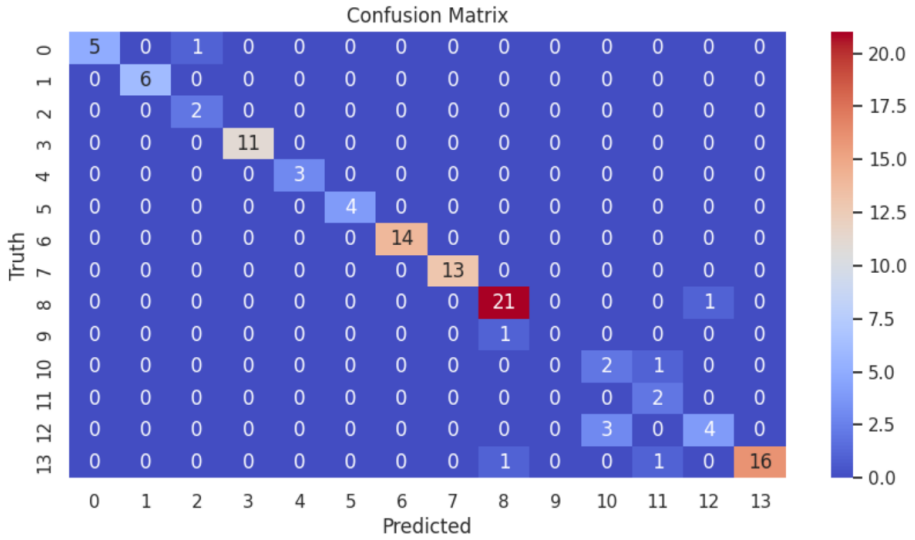

Fig. No. 6: Confusion Matrix of Fertilizer Prediction Model.

The confusion matrix evaluates the model's performance by comparing actual and predicted fertilizer classifications. The diagonal cells represent correct predictions, while off-diagonal cells indicate misclassifications. The color intensity shows the frequency of predictions, with darker colors highlighting areas of higher classification accuracy or error. This matrix helps

assess the model's strengths and areas for improvement.

This image is a decision tree visualization showing how the model classifies different fertilizers based on input features. Each node represents a decision rule, splitting data into branches that lead to a final fertilizer prediction. The tree structure helps in understanding how different soil and crop parameters influence the fertilizer recommendation.

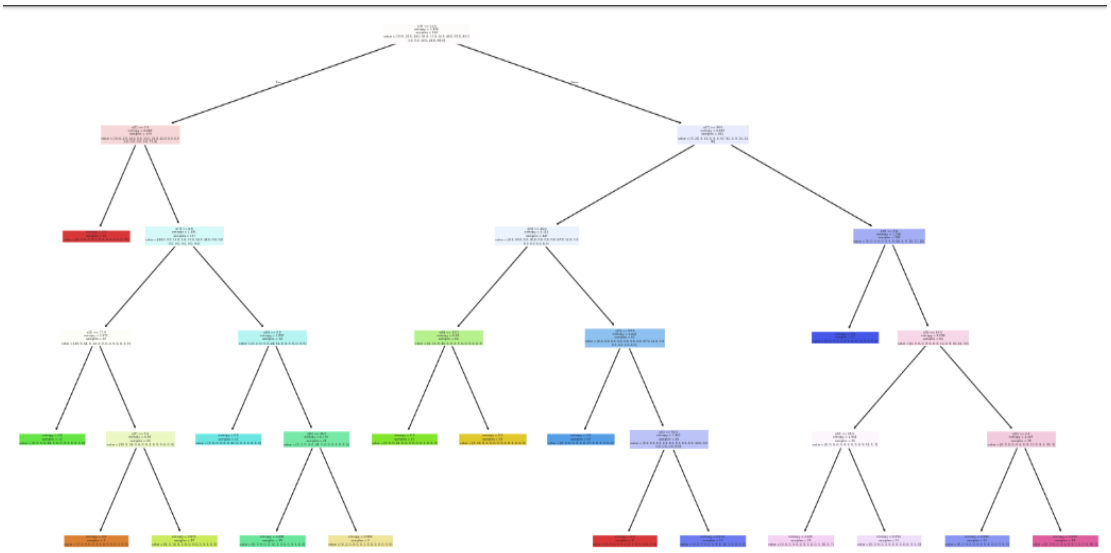

Fig. No. 7: Decision Tree Model for Fertilizer Classification.

```
#accuracy of the model
acc.append(metrics.accuracy_score(y_test,y_pred)*100)
model.append('Decision Tree')

print('Accuracy of the model :',metrics.accuracy_score(y_test,y_pred)*100)
```

```
Accuracy of the model : 90.09009009009009
```

```
#accuracy of the model
acc.append(metrics.accuracy_score(y_test,y_pred)*100)
model.append('Decision Tree')

print('Accuracy of the model :',metrics.accuracy_score(y_test,y_pred)*100)
```

```
Accuracy of the model : 94.18604651162791
```

The accuracy of the model predicted is 94.18%

V. CONCLUSION

The accuracy of the Previous model is 90.96%, as calculated. This indicates that the model correctly predicts the target variable for approximately 90.96% of the test data. However, when compared to another model, such as the Decision Tree Model, which has an accuracy

of 94%, we observe a performance difference of 2.04%.

The 2.04% accuracy gap highlights the advantage of using methods like Decision Tree Model, which is used to provide a more robust and stable prediction. The Decision Tree model is a good choice for quick interpretations and lower computational costs.

REFERENCE

1. Amelia Devi Putri A Riyanto, Diana Parnitas're, and Chastine Faatiha, "A Systematic Review on Semantic Relabeling for Information Extraction in Low-Resource Data". IEEE Transaction, April 2024; 12: 57917-57946. DOI: 10.1109/ACCESS.2024.3392370
2. Rong Xin Liaoyuan Lulea Shi, and Shuri Tan, "Research on Multi-Label Semi-Supervised Learning Algorithm Based on Dual selection Criteria". IEEE Transaction, Feb2024; 12: 31357-31365. DOI: 10.1109/ACCESS.2024.3369919
3. Raed Alizadeh, Mazen Aloud, Haneen Alzoubi, "The significance of capturing the correlations among labels in multi-label classification: An investigative study". ResearchGate, 2023; 1: 060005-(1-9), DOI: 10.1063/5.0177340
4. Raed Alizadeh, Ghassan Samara, Sattam Agmatine, Mohammad Hassan, Mohammad ALADI and Hasan Mansur, "Multi-Label Classification Based on Associations". MDPI, 2023; 13: 1-16. DOI: 10.3390/app13085081
5. Karelian Medeiros Ovidio Vale, Arthur Costa Gorgonio, Flavius Da Luz E Gorgonio, Anne Magaly De Paula Canuto, "An Efficient Approach to Select Instances in Self-Training and Co-Training Semi-Supervised Methods". IEEE transaction, 2022; 10: 7254-727. DOI: 10.1109/ACCESS.2021.3138682
6. Rishita Ray, Tanmayi Nandan, Lahari Aneke'll Anil Kumar, "A New Combined Model with Reduced Label Dependency for Malware Classification". ResearchGate, 2021; 2: 23-32. DOI: 10.2991/ahis.k.210913.004
7. Teodor Fredriksson, David Issa Mattos, Jan Bosch and Helena Holmstrom Olsson, "Data Labelling: An Empirical Investigation into Industrial Challenges and Mitigation Strategies". ResearchGate, 2020; 1-16, DOI: 10.1007/978-3-030-64148-1_13
8. Raed Alizadeh, Farzana Kabir Ahmad, Mohamad Farhan Label Ranking Method Based on Positive Class Correlations". ResearchGate, Dec2020; 06: 377-399, DOI: 10.5455/jjcit.71-1592597688
9. Farzana Ahmad, and Mohamad Mohsin School of Computing, University Utara Malaysia, Malaysia, "Multi Label Ranking Based on Positive Pairwise Correlations

Among Labels". Research Gate, July 2019; 1-11. DOI: 10.34028/fajita/17/4/2

10. Muhammad Usama, Junaid Qadir, Aunn Raza, Hunain Arif, Kok-Lim Alvin Yau, Yehia Elkhatib, Amir Hussain, And Ala Al-Fuqaha, "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges". IEEE Transaction, June 2019; 7: 10.1109/ACCESS.2019.2916648 No: 65579-65615.