Comparative analysis of data exploration techniques

Tomás Brás 112665
tomasbras@ua.pt

***Abstract*—Machine Learning has emerged as a fundamental tool for analyzing and extracting knowledge from large datasets across various domains, including healthcare, finance, and industry. This paper presents an evaluation of different ML techniques, categorizing them into Supervised and Unsupervised Learning models. The study explores key regression and classification algorithms for supervised learning, as well as clustering and dimensionality reduction techniques for unsupervised learning.**

The analysis identifies the most complete algorithms within each learning paradigm by evaluating their performance across multiple key metrics. In Supervised Learning for regression tasks, Gradient Boosting Regression emerges as the most robust algorithm, consistently excelling in all evaluation criteria. For classification, Neural Networks and Support Vector Machines (SVM) stand out for their balanced performance in accuracy, F1-Score, and AUC-ROC, making them highly reliable in varied scenarios. In Unsupervised Learning, K-Means proves to be the most well-rounded clustering algorithm, while PCA and ICA dominate in dimensionality reduction. These results highlight the importance of selecting versatile algorithms that perform strongly across multiple dimensions of evaluation.

The study also includes real-world examples, such as election prediction using SVM and defect detection in metals using PCA and ICA, to illustrate the practical application of these techniques. Results demonstrate that algorithm performance varies depending on the dataset and evaluation criteria, reinforcing the importance of metric-aware algorithm selection. The findings contribute to a better understanding of the strengths and limitations of each method, offering guidance for more effective data-driven decision-making.

***Index Terms*—Machine Learning, Techniques, Efficiency, Information Extraction, Supervised, Unsupervised, Data Analysis.**

I. Introduction

In recent decades, the amount of information associated with our surroundings has increased exponentially as a result of technological advancements. As a consequence of this radical growth, it has become possible to extract large volumes of data that, through artificial intelligence, can be used to automate processes and drive innovation in various fields, such as medicine, education, finance, and industry. Artificial intelligence enables systems to enhance their functions in a way that mimics human capabilities, making it possible to solve problems and learn from experience and received data. This advancement allows tasks that previously required human intervention to be performed autonomously, increasing efficiency and precision in technological processes.

A. Machine Learning

The branch of artificial intelligence responsible for analyzing and processing large amounts of data is Machine Learning. The goal of Machine Learning is to analyze vast datasets, identify patterns, and ensure that computational systems can make decisions without needing to be manually adapted to the shared information.

The learning process is divided into two main phases: training and testing. [1] The training phase begins with data analysis and study; however, before this analysis, data preprocessing is required to remove inconsistencies such as missing values, duplicate data, and contradictory entries. Before training begins, it is essential to select the appropriate learning model.

Next, a sample of the data is used to train the model. Through data analysis and algorithm formulation, the model learns to recognize patterns and automate decision-making. After training, the model undergoes a testing phase, where processed data is used to evaluate its **accuracy and generalization capability**. The trained model is used to analyze new data and generate predictions. This process ensures that the model can be applied to new datasets and continues to **improve its efficiency** as it receives additional data over time.

B. Types of Machine Learning

Depending on the nature of the data used to train the model, we need to choose the most appropriate type of Machine Learning to apply. Machine Learning models can be classified as follows:

- **Supervised Learning**
- **Unsupervised Learning**
- **Other Learning Types:**
  - **Semi-Supervised Learning**
  - **Reinforcement Learning**
  - **Active Learning**
  - **Online Learning**

1) Supervised Learning: Supervised learning is a type of Machine Learning where the model learns from **labeled data**. This means that each data input has an associated label representing the expected output. [1] The goal is to associate x (input data) with t (expected output) and, based on this, identify patterns in the input data, allowing the model to make predictions using the acquired knowledge. The more labeled examples provided to the model, the greater its ability to recognize patterns and generalize to new data.

This type of Machine Learning aims to solve two main problems: **Classification**, when the objective is to predict categories or classes (e.g., determining whether an email is spam or not). **Regression**, when the objective is to predict continuous values (e.g., forecasting house temperature).

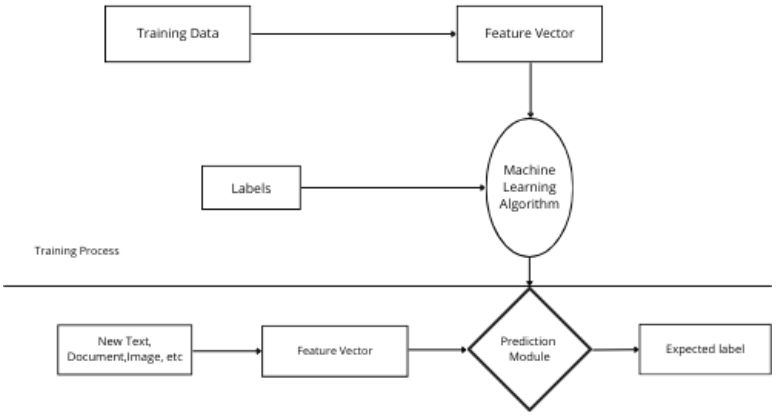

Fig. 1: Supervised learning model

2) Unsupervised Learning: Similar to supervised learning, unsupervised learning also processes large volumes of data. However, these data are **not labeled**. The goal is to analyze the data and identify similarities between them, even without prior knowledge of their categories. [1]

This approach is applied to problems such as: **Clustering**, where the objective is to group data points with common characteristics without requiring predefined labels. **Dimensionality Reduction**, which reduces the number of variables in a dataset while preserving as much information as possible (e.g., image compression).

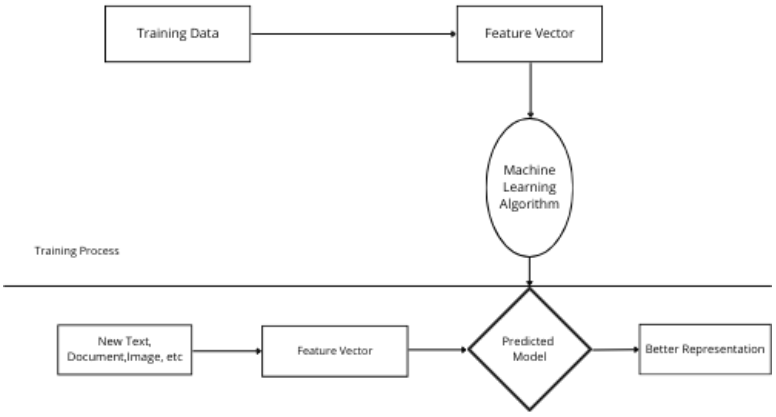

Fig. 2: Unsupervised learning model

3) Other Learning Types: Semi-Supervised Learning analyzes a mixed dataset, consisting of a small set of labeled data and a large set of unlabeled data. Reinforcement Learning learns through interaction with an environment. Active Learning selects the examples it wants to learn, reducing the need for large volumes of labeled data. Online Learning continuously adapts to new data in real time, without needing to restart training from scratch.

C. Techniques Used in Supervised and Unsupervised Learning

For different learning models, a wide range of algorithms is available. Figure 3 presents the main algorithms used in supervised and unsupervised learning [2] [3] [4] [5] [6] [7] [8]. Additionally, there are algorithms for **Reinforcement Learning**, such as **Q-Learning** and **Deep Q-Networks (DQN)**. For **semi-supervised learning**, key techniques include **Self-training, Co-training, and Generative Adversarial Networks (GANs)**.

These algorithms play a fundamental role in data analysis, allowing the extraction of knowledge from large datasets and enabling the development of **autonomous and efficient computational systems**.

Thus, the next step of this study will be to **comparatively analyze these algorithms**, investigating their advantages, limitations, and relationship with the dataset. This approach will provide a better understanding of their role within each Machine Learning model.

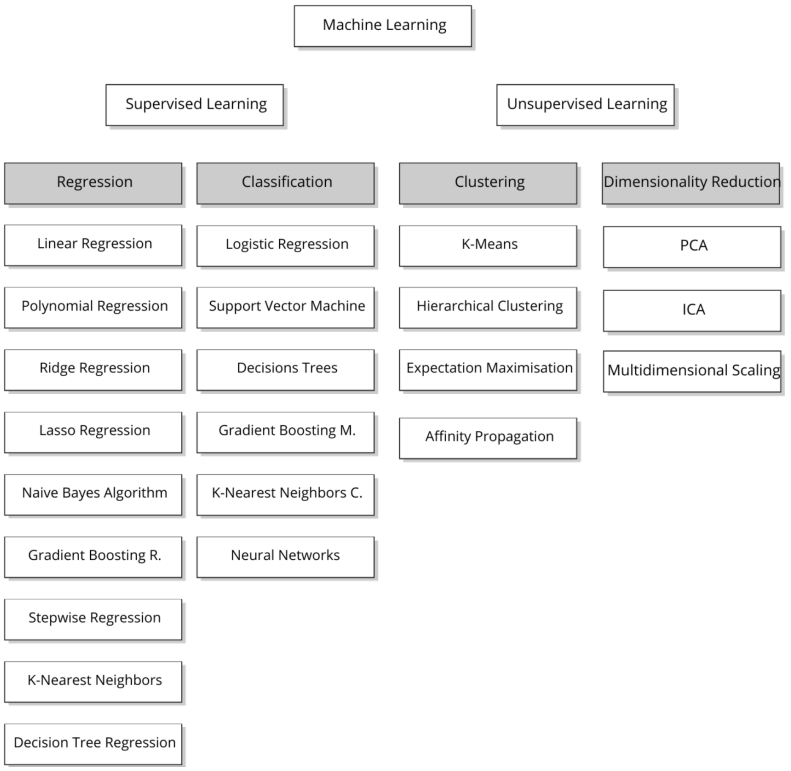

Fig. 3: Machine Learning Tree

II. Methodology

A. Algorithm Performance Evaluation Criteria

The algorithms identified in Figure 3 are categorized according to the learning model. To conduct a comparative analysis, various **performance metrics** were used. These metrics assess the accuracy and quality of a model, ensuring that it makes **coherent and high-performance predictions** based on real data. In Figures 4, 5, and 6, the performance of different algorithms is identified according to the evaluation metrics

used. The graphical representation follows the following classification criteria:

- ✪ indicates that the algorithm had a **Very Good** performance for that metric.
- ✓represents a **Good** performance, but not necessarily the best among the evaluated methods.
- An empty field indicates that the algorithm had a **Poor** performance or that the metric **does not apply** to that method.

1) Performance Metrics for Supervised Learning Algorithms: For the Supervised Learning model, we consider two categories: **Regression and Classification**. Some algorithms work for both problems, such as **Decision Trees, Neural Networks, SVM, Gradient Boosting, and K-Nearest Neighbors**.

2) Performance Metrics for Regression Algorithms: For **regression models**, the following performance metrics were used [9]:

- **Mean Squared Error (MSE)**: Measures the average squared errors between actual and predicted values. The lower the MSE, the better the model, as it indicates a lower average error.
- **Root Mean Squared Error (RMSE)**: The square root of the MSE, maintaining the same unit as the original values and penalizing large errors more significantly.
- **R^2 (R-Squared)**: Quantifies the proportion of data variance explained by the model. A value close to 1 indicates a good fit, while lower values suggest that the model does not adequately represent the data.
- **Mean Absolute Error (MAE)**: Calculates the average absolute error between predicted and actual values, without considering the error direction. This metric is useful for evaluating prediction accuracy in a more intuitive way.
- **Mean Absolute Percentage Error (MAPE)**: Measures the average percentage error relative to actual values, making it useful for interpreting the relative error across different data scales.

Ref. No	Methods	MSE	RMSE	R^2	MAE	MAPE
[3] [8]	Linear Regression	✓	✓	✓		
[8]	Polynomial Regression			✪		
[8]	Ridge Regression	✪	✪	✪	✓	✓
[8]	Lasso Regression	✪	✪	✪	✓	✓
[6]	Naïve Bayes Algorithm					
[8]	Gradient Boosting R.	✪	✪	✪	✪	✪
[8]	Stepwise Regression	✓	✓	✓	✓	
[6]	K-Nearest Neighbors				✓	✓
[3] [6]	Decision Tree Regression				✓	✓

TABLE I: Fig. 4 - Comparison Table of Performance Analysis for Supervised Learning Machine Regression Algorithms

3) Performance Metrics for Classification Algorithms:

- **Accuracy**: Measures the proportion of correct predictions in relation to the total number of samples. Although higher accuracy generally indicates better overall classification performance, **accuracy has lost credibility as a performance metric, particularly in scenarios involving imbalanced datasets** [10]
- **Sensitivity**: Evaluates the model's ability to correctly identify instances of the positive class. High sensitivity means that the model minimizes false negatives, ensuring that fewer positive cases are mistakenly classified as negative.
- **Efficiency**: Relates to computational time and the resources required for model training and inference. More efficient models achieve good performance with lower computational cost, making them ideal for real-time applications or hardware-constrained environments.
- **F1-Score**: Combines precision and sensitivity into a single metric by calculating the harmonic mean between both. This metric is particularly useful when there is an imbalance between classes, as it prevents a model from achieving good results simply by favoring the majority class due to data distribution.
- **AUC-ROC**: Measures the model's ability to distinguish between classes. The higher the AUC, the better the model separates positive and negative classes, making it a fundamental metric for evaluating probabilistic classifiers and their discrimination capability.

Ref. No	Methods	Accuracy	Sensitivity	Efficiency	F1-Score	AUC-ROC
[6][8]	Logistic Regression	✓		✪	✓	✓
[6][8][11]	Support Vector Machine	✪	✓		✪	✪
[6]	Decision Trees	✓	✓	✓	✓	
[8]	Gradient Boosting M.	✪	✪		✪	✪
[6]	K-Nearest Neighbors C.		✪		✓	✓
[6][10]	Neural Networks	✪	✪		✪	✪

TABLE II: Fig. 5 - Comparison Table of Performance Analysis for Supervised Learning Machine Classification Algorithms

4) Performance Metrics for Unsupervised Learning Algorithms: For the Unsupervised Learning model, we have two categories: **Clustering** and **Dimensionality Reduction**. In this model, as presented in Fig. 6, we include some algorithms that address clustering problems, while the last three correspond to dimensionality reduction problems.

- **Silhouette Coefficient**: Measures the quality of clusters by calculating how similar a point is to its own cluster compared to other clusters. Higher values indicate better clustering, reflecting greater internal coherence within clusters.
- **Davies-Bouldin Index**: Evaluates the compactness and separation of clusters. The lower the index, the better the separation between clusters, indicating that the formed groups are more distinct and well-defined.
- **Calinski-Harabasz Index**: Analyzes cluster dispersion by comparing the variance between and within clusters. Higher values indicate better-structured clustering, where points within each cluster are closer to one another, and clusters are well separated.

- **Explained Variance Ratio**: Used in dimensionality reduction algorithms, such as PCA, to measure the amount of variance retained after transformation. Higher values indicate less information loss, ensuring that the transformation preserves the structure of the original data.
- **Trustworthiness Score**: Assesses whether the local structure of the data is preserved after dimensionality reduction. A high value indicates that close neighbors in the original data remain close in the new dimensional space, ensuring that the transformation does not distort important relationships between data points.

Ref. No	Methods	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz Index	Explained Variance Ratio	Trustworthiness Score
[1][3][4]	K-Means	✔	✔	✔		
[1][4]	Hierarchical Clustering	✔	✔	✔		
[4]	Expectation Maximisation	✓	✔	✔		
[7]	Affinity Propagation	✓				
[3]	Expectation-Maximization	✔		✔		
[4][11]	PCA				✔	✓
[4]	Multidimensional Scaling					✔
[4][11]	ICA					✔

TABLE III: Fig. 6 - Comparison Table of Performance Analysis for Unsupervised Learning Algorithms

B. Algorithms and Data Analysis

Although most algorithms are designed for predictive tasks, their performance can vary significantly depending on the evaluation metric considered. Therefore, when choosing an algorithm within a learning model (e.g., classification or regression), it is advisable to select the one that shows the **best results in the most relevant performance metrics for the specific problem**. This ensures more accurate and reliable predictions, especially when applied to real-world data.

1) Example 1: Predicting Election Results with Supervised Learning: In the study conducted by **Malhar Anjaria and Ram Mohana Reddy Guddeti** [11], the goal was to predict **election results** using **supervised learning**, applied to the **2012 U.S. presidential elections** and the **2013 Karnataka Assembly elections in India**. The approach involved classifying tweets as **positive or negative** to determine public sentiment towards candidates.

The study compared four well-known supervised learning algorithms for sentiment analysis and text classification. **Naïve Bayes (NB)** is a probabilistic model effective for large datasets such as tweets. It assumes independence between words in the text, which can be a limitation, but it performs well in practical text categorization tasks. **Maximum Entropy (MaxEnt)** does not assume independence among features, allowing for a more precise adjustment of class probabilities. **Support Vector Machines (SVM)** was used to find an optimal hyperplane that separates positive and negative tweets, achieving the highest accuracy in the study, making it the best-performing classifier for predicting election results. **Artificial Neural Networks (ANN)** were capable of capturing complex relationships and non-linear patterns in tweets. Although they provided good results, their high computational cost made them less efficient for this application.

Additionally, an **Unsupervised Learning algorithm (PCA - Principal Component Analysis)** was used as a preprocessing step to optimize the performance of SVM. PCA was applied to reduce data dimensionality and eliminate redundant variables, minimizing the risk of overfitting and making the model more generalizable to new tweets. The results demonstrated that **SVM outperformed the other models**, achieving the highest accuracy in both election datasets: **88% for the 2012 U.S. Presidential Elections** and **58% for the Karnataka Assembly Elections 2013**. This indicates that SVM is an effective method for sentiment analysis in political forecasting, especially when combined with dimensionality reduction techniques like PCA. [11]

2) Example 2: Detecting Microscopic Defects in Metals with Unsupervised Learning: The study conducted by **Xin Zhang, Jafar Saniie, Sasan Bakhtiari, and Alexander Heifetz** [12] focused on using **Unsupervised Learning** to detect **microscopic defects in metals** through **Pulsed Infrared Thermography (PIT)** [12]. The objective was to identify internal porosities in metallic components used in **nuclear reactors**, ensuring structural integrity.

Four Unsupervised Learning algorithms were compared, with emphasis on **PCA (Principal Component Analysis)** and **ICA (Independent Component Analysis)**. PCA was used to reduce the dimensionality of thermal data, extracting the most relevant thermal patterns and removing unwanted noise. It maximized data variance and identified principal components, allowing for a more efficient reconstruction of thermal images. ICA, on the other hand, was applied to separate mixed thermal signals into independent components, enabling the identification of distinct thermal patterns associated with structural defects [12]. ICA proved effective in segmenting defective areas more accurately.

The study concluded that ICA outperformed PCA in detecting thermal defects, as it extracted more relevant thermal patterns. However, this came at a higher computational cost due to the iterative optimization required to find independent components. PCA, on the other hand, was faster and computationally more efficient, making it a viable choice when speed is prioritized.

III. Discussion

The evaluation of different Machine Learning models based on performance metrics provides clear insights into the top-performing algorithms across Supervised and Unsupervised Learning categories. By analyzing the performance tables, it is possible to identify the most complete algorithms for each task. For Supervised Learning, **Gradient Boosting Regression** stands out by consistently excelling across all metrics, making it the most robust option for complex prediction tasks. **Lasso**

and Ridge Regression also perform well, especially in high-dimensional scenarios prone to overfitting. In classification tasks, **Neural Networks** and **Gradient Boosting Machines** offer high accuracy and flexibility, while **Support Vector Machines** are particularly effective for binary classification with clear class separation. In the Unsupervised Learning category, **K-Means** proves to be the most complete clustering algorithm for well-separated data, while **Hierarchical Clustering** is more suitable for data with nested group structures. For dimensionality reduction, **PCA** excels in preserving global variance, whereas **ICA** is preferable when local structure is more relevant. Combining Supervised and Unsupervised techniques such as using PCA before training an SVM can further enhance performance, demonstrating the value of hybrid approaches. Finally, it's important to emphasize that the **"best"** algorithm depends on the evaluation metric used. An algorithm with excellent RMSE may not perform equally well in terms of R^2. Therefore, selecting the appropriate model must be aligned with the specific objectives and evaluation criteria of the task.

IV. Critical Thinking

Through this work, I acquired fundamental knowledge that could shape my future in the field of data analysis and Machine Learning algorithms. Initially having limited background in this area, I was able to gain a clearer understanding of when and how to apply different Machine Learning models, particularly **Supervised** and **Unsupervised Learning**.

I learned that **Supervised Learning** is appropriate for prediction tasks, where data is labeled and the expected outcome is known — for example, predicting prices, classifying emails as spam, or diagnosing diseases. In contrast, **Unsupervised Learning** is used when data is unlabeled, making it suitable for discovering hidden patterns, such as customer segmentation or dimensionality reduction in complex datasets like images or biomedical signals.

Moreover, I realized that choosing the most appropriate algorithm depends not only on the model type but also on the **data characteristics, the problem structure, and the evaluation metrics** aligned with the goal. For instance, one algorithm may perform better in terms of *RMSE*, while another achieves a higher R^2. Therefore, selecting the metric that best fits the problem is crucial for making informed decisions.

Throughout the analysis, important questions arose regarding the use of performance metrics. I consider them essential tools, as they help identify the strengths of each algorithm. One specific concern involved the widespread use of **accuracy**. Although it is one of the most popular metrics, I learned that it can be **misleading in imbalanced datasets** — such as in fraud detection, where a model could achieve 99% accuracy simply by always predicting the majority class. Nevertheless, I decided to include accuracy in the study to observe which algorithms are more affected by such bias and to become more aware of its limitations.

V. Conclusion

This study highlights how different **Machine Learning** approaches, both **Supervised and Unsupervised Learning**, can be applied to diverse data analysis problems. This work underscores the importance of selecting appropriate algorithms based on specific problem domains, ensuring efficient, scalable, and accurate solutions in **data science, industry, and artificial intelligence applications**. In conclusion, this work provided me with a practical and critical perspective on the foundations of Machine Learning.

Acknowledgment

This paper was supported by AI tools, specifically *elicit.com*, which assisted in literature review and citation analysis by providing relevant academic sources.

References

[1] R. Reddy and G. K. Shyam, "Analysis through machine learning techniques: A survey," in *2018 IEEE International Conference on Green Computing and Internet of Things (ICGCIoT)*. IEEE, 2018, pp. 542–546.

[2] U. Muqtar, A. Ameen, and S. Raziuddin, "Opinion mining on twitter data using unsupervised learning technique," *International Journal of Computer Applications*, vol. 148, no. 12, 2016.

[3] R. Soni and K. J. Mathai, "Effective sentiment analysis of a launched product using clustering and decision trees," *International Journal of Innovative Research in Computer and Communication Engineering*, 2017.

[4] M. Usama, J. Qadir, A. Raza, H. Arif, K.-L. A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, "Unsupervised machine learning for networking: Techniques, applications and research challenges," *IEEE Access*, vol. 7, pp. 65 579–65 617, 2019.

[5] R. Saravanan and P. Sujatha, "A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification," in *Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018)*. IEEE, 2018, pp. 945–949.

[6] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *Proceedings of the 10th INDIACom; INDIACom-2016; IEEE Conference ID: 37465, 2016 3rd International Conference on "Computing for Sustainable Global Development"*. New Delhi, India: IEEE, 2016, pp. 1310–1315. [Online]. Available: https://ieeexplore.ieee.org/document/1310978

[7] B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu, "Learning vector quantization for (dis-) similarities," *Neurocomputing*, vol. 131, pp. 43–51, 2014.

[8] A. Abdulhafedh, "Comparison between common statistical modeling techniques used in research, including: Discriminant analysis vs logistic regression, ridge regression vs lasso, and decision tree vs random forest," *Open Access Library Journal*, vol. 9, p. e8414, 2022. [Online]. Available: https://doi.org/10.4236/oalib.1108414

[9] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. ACM, 2006, pp. 161–168. [Online]. Available: https://doi.org/10.1145/1143844.1143865

[10] I. Syed and V. Lokhande, "An overview of the supervised machine learning," *International Research Journal of Modernization in Engineering, Technology and Science*, vol. 6, no. 3, pp. 6355–6360, 2024, peer-Reviewed, Open Access, Fully Refereed International Journal. [Online]. Available: https://www.irjmets.com

[11] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of twitter data using supervised learning," *Journal of Computer Science*, 2015.

[12] X. Zhang, J. Saniie, S. Bakhtiari, and A. Heifetz, "Unsupervised learning for detection of defects in pulsed infrared thermography of metals," in *2022 IEEE International Conference on Electro Information Technology (eIT)*. IEEE, 2022, pp. 1–6.