**Indian Journal of Modern Research and Reviews**

**This Journal is a member of the *'Committee on Publication Ethics'***

**Online ISSN:**2584-184X

**Research Paper**

# A Systematic Review of Machine Learning Algorithms for Classification: General Approaches and Environmental Applications

**Nomsa C. C. Kamgwira [1], Shalu Gupta [2*]**

[1] Student, Department of Computer Applications, Guru Kashi University, Talwandi Sabo, Punjab, India

[2] Associate Professor, Department of Computer Applications, Guru Kashi University, Talwandi Sabo, Punjab, India

**Corresponding Author:** *Shalu Gupta

**DOI:** https://doi.org/10.5281/zenodo.17722101

## ABSTRACT

The field of Machine Learning has seen rapid advancement from 2022 to 2025 due to more cutting-edge computational tools, hybrid models and improved specified techniques. This review has been written to assess the widely used classification algorithms, such as traditional, ensemble-based and deep learning. It evaluates their performance in practical applications. The search covered five open learning databases, which are: Google Scholar, Semantic Scholar, arXiv, DOAJ and ResearchGate. 57 studies published between 2022 and 2025 met the selection criteria. Findings show that Random Forests and XGBoost are effective for structured datasets, and CNNs and transformers are more suitable for unstructured datasets. Hybrid deep learning ensembles are more stable as they can capture spatial and temporal patterns. This review provides a summary of the results, including a comparison table and an outline of areas that require further work.

**Manuscript Info.**

- ✓ **ISSN No:** 2584- 184X
- ✓ **Received:** 08-9-2025
- ✓ **Accepted:** 29-10-2025
- ✓ **Published:** 26-11-2025
- ✓ **MRR:3(11):**2025;41-47
- ✓ **©2025, All Rights Reserved.**
- ✓ **Peer Review Process:** Yes
- ✓ **Plagiarism Checked:** Yes

**How To Cite this Article**

Kamgwira NCC, Gupta S. A systematic review of machine learning algorithms for classification: general approaches and environmental applications. Indian J Mod Res Rev. 2025;3(11):41-47.

**KEYWORDS:** Machine learning; Classification algorithms; Ensemble learning; Deep learning; Transformers; Environmental classification; Systematic review; Open-access databases.

## 1. INTRODUCTION

Machine learning (ML) classification algorithms play an important role in data-driven decision-making across fields such as healthcare, finance, cybersecurity, agriculture, and environmental science. Between 2022 and 2025, research on classification accelerated due to larger datasets, improved computational tools, and expanded open-source ML libraries. Recent algorithms, mainly transformer-based models, have enhanced the abilities of traditional classification approaches.

Even though there are a lot of review papers out there, the focus is on a single field or specific related algorithms. Some studies combine general algorithm performance with a deeper understanding of real-world applications.

Environmental classification has become increasingly relevant with climate change (rainfall prediction, pollution categorisation, and weather-related risk assessment), yet research in this area is still scattered [1,2].

**The objective of the review is to address these gaps by:**

1. Presenting an updated (2022–2025) systematic review of commonly used ML classification algorithms

2. Comparing performance trends across structured, image, text, and sensor-based datasets
3. Providing environmental classification as a practical example
4. Adopting open research practices that promote the sharing of data, methods, and results
5. Presenting tables and summarised findings for researchers and postgraduate students

**ML classification algorithms fall into three groups:**

### 1.1 Traditional Classification Algorithms

Common traditional algorithms are Logistic Regression and Support Vector Machines (SVM), Naïve Bayes, k-Nearest Neighbours (k-NN), and Decision Trees. They work well with structured data and small-to-medium datasets. Current research explores improvements such as kernelised SVMs, providing improved clarity and understanding [6].

### 1.2 Ensemble-Based Algorithms

Ensemble models are more powerful and are created by pooling the results from several individual predictive models. Common models are Random Forest, AdaBoost, Gradient Boosting, CatBoost, and XGBoost. As of 2022, XGBoost and CatBoost gained traction for handling common issues efficiently, like handling missing data, categorical variables, and class imbalance [7].

### 1.3 Deep Learning and Transformer-Based Algorithms

Deep learning approaches, specifically CNNs, RNNs, LSTMs, and transformers, are broadly applied to unstructured data, which encompasses images, audio, and text [8].

Transformers were first developed for Natural Language Processing (NLP). However, they are now used in time-series and environmental applications, showing strong performance in recognising complex patterns [1].

### 1.4 Environmental Classification as a Case Example

Environmental datasets typically include:

- Nonlinear patterns
- High temporal variability
- Spatial dependencies
- Noisy or missing observations

These features make environmental classification a suitable test case for algorithm robustness. Recent studies demonstrate that hybrid CNN–LSTM models, XGBoost-based feature selection pipelines, and transformer architectures perform well in classifying rainfall intensity, air quality, and drought severity [10].

### 1.5 Purpose of This Review

This study analytically reviews machine learning classification algorithms using open-access literature available from 2022 to 2025, evaluating standard, widely-used machine learning methods in classifying environmental data. The review aims to clarify:

- Changing dynamics in classification trends
- Effective hybrids within the field
- Current limitations and areas for future study

## 2. METHODOLOGY

This review followed a protocol modelled on the PRISMA 2020 guidelines, which is a widely recognised framework designed to improve transparency of systematic reviews. The methodology makes sure that the research process is clear and can be reproduced. This allows other researchers to replicate the same steps and confirm the findings.

The WeatherAUS dataset is used to predict weather events in Australia as it presents a significant class imbalance. There are more instances of ‘No Rain’ than ‘Rain’. This imbalance raises difficulties for machine learning models, which tend to prefer the majority class. This results in exaggerated accuracy scores while underperforming in classifying actual rain events. As a result, the F1 score, which balances precision and recall, reduces for the minority class [4].

This problem is well-known in environmental prediction research, though rare but critical events like rainfall are occasional and highly variable, making them harder to model effectively. To resolve imbalance effects, standard preprocessing techniques such as stratified splitting, normalisation were applied. However, the imbalance remains a key factor influencing the experimental outcomes and must be taken into consideration when interpreting classifier performance [5].

### 2.1 Search Strategy

The literature search was performed exclusively on open-access platforms. The following databases were used:

- Google Scholar: primary search engine
- Semantic Scholar: links to open-access versions
- arXiv: preprints in machine learning and environmental modelling
- Directory of Open Access Journals (DOAJ): peer-reviewed open-access journals
- ResearchGate: preprints or accepted manuscripts uploaded by authors

**Search keyword**

Keywords and Boolean operators included:

- “Machine learning classification”
- “Supervised learning models”
- “Classification algorithms review”
- “Deep learning classifier”
- “Transformer classification model”
- “Environmental classification” OR “weather classification”
- “Rainfall classification” AND “machine learning”

**Search timeframe**

Only studies published between January 2022 and January 2025 were considered.

### 2.2 Inclusion Criteria

Studies were included if they met all of the following:

- Published 2022–2025
- Free PDFs and open-access preprint
- Applied and discussed machine learning classification algorithms

- Reported performance metrics such as accuracy, F1-score, precision, recall, or AUC
- Written in English
- Peer-reviewed OR preprint from arXiv with established author credibility
- For a domain example: studies applying ML to environmental classification tasks

### 2.3 Exclusion Criteria

Studies were excluded if they:

- Required paid access (e.g., publisher paywalls)
- Focused solely on regression, clustering, or reinforcement learning
- Did not include any classification model
- Lacked methodological detail
- Provided no quantitative results
- Duplicate entries
- Non-English or inaccessible manuscripts

### 2.4 Study Selection Process (PRISMA Description)

A total of 265 records were found across all free-access databases. After removing duplicates, checking titles and abstracts, 121 studies were left. A detailed assessment removed articles lacking relevance (e.g., missing metrics, unclear methodology). This left [8]:

- Picked studies: 57 (2022–2025)

### PRISMA-Style Flow Description

### Identification

- Records found through open-access databases: 265
- Duplicates removed: 56

### Screening

- Records screened via title/abstract: 209
- Records excluded for irrelevance: 88

### Eligibility

- Detailed articles assessed: 121
- Detailed excluded (inaccessible, not classification, lacked metrics): 64

### Included

- Studies included in the final review: 57

### 2.5 Data Extraction

For each included paper, the following information was reviewed:

- Publication year
- Algorithm types used
- Dataset characteristics
- Field of study (general vs. environmental)
- Evaluation metrics
- Model comparison results
- Key contributions or innovations
- Limitations

Extraction was done manually to avoid bias and ensure accuracy.

### 2.6 Quality Assessment

The quality of included studies was evaluated according to several key criteria:

- Clarity of data description
- Availability of the dataset
- Explanation of preprocessing steps
- Validation technique
- Reporting of multiple metrics
  Studies scoring below 50% on quality indicators were excluded.

## 3. RESULTS

This section shows the findings from the 57 included studies and combines an experimental section conducted by the researcher to add to the literature-based findings. This experiment is to check commonly used classification algorithms on a public dataset to measure their performance.

### 3.1 Literature-Based Results: General Classification Algorithms

Analysis of the studies used (2022–2025) shows clear trends:

- Ensemble models (XGBoost, CatBoost, Random Forest) always outperform traditional algorithms on structured datasets.
- CNNs are dominant for image- and sensor-based classification tasks.
- Transformers give the best results on text, satellite imagery, and time-series environmental classification.

Hybrid models combining deep learning with ensemble approaches emerged as the highest-performing category in complex environmental tasks.

**Table 1:** Summary of General Classification Algorithm Performance (2022–2025)

| Algorithm | Typical Accuracy Range | Best Data Type | Key Strength | Common Limitation | Example Open-Access Study |
|---|---|---|---|---|---|
| Logistic Regression | 70–85% | Tabular | Interpretable | Limited to linear trends | Khan et. al. (2022) [10] |
| IGWO SVM | 80–98.75% | Tabular/Image | Good for small datasets | High training cost | Shen et. al. (2023) [11] |
| Random Forest | 85–96% | Tabular | Robust | Memory-heavy | Ahmed et al. [12] |
| XGBoost | 88–98% | Tabular | Handles missing data | Requires tuning | Wu & Zhang (2023) [13] |
| CatBoost | 87–97% | Categorical | No encoding needed | Slower | Li et al. [14] |
| CNN | 90–99% | Image/Sensor | Best for spatial data | Needs large data | Singh et al. [15] |
| LSTM/GRU | 83–95% | Time Series | Models' long sequences | Slow training | Ogundele et al. [16] |
| Transformer | 92–99% | Text/Image/TS | State-of-the-art | High cost | Chen et al. [17] |

### 3.2 Literature-Based Results: Environmental Classification

Environmental applications included rainfall prediction, flood risk classification, drought severity analysis, and air quality categorisation.

**Key developments:**

- CNN–LSTM hybrids excel in tasks that have spatial and temporal dependencies.
- XGBoost performs best for air-quality sensor data due to the tabular structure.
- Transformers (especially time-series variants) are emerging leaders in rainfall and pollution classification.

**Table 2:** Environmental Classification Algorithms and Findings (2022–2025)

| Study | Task | Best Performing Model | Metric (Reported) | Key Insight |
|---|---|---|---|---|
| Wang et al., 2022 [18] | Rainfall event classification | CNN–LSTM | 93.6% accuracy | Hybrid deep models capture spatial + temporal rainfall patterns effectively. |
| Cheng et al., 2023 [19] | Air quality level classification | XGBoost | 95–97% accuracy | Gradient boosting handles sensor noise and missing values well |
| Huang et al., 2024 [20] | Flood susceptibility classification | Transformer-based encoder | AUC = 0.92 | Transformer attention improves long-term hydrological dependency modelling |
| Rahman et al., 2023 [21] | Land cover & satellite image classification | CNN | 98% accuracy | CNNs best extract spatial spectral patterns in remote sensing images |
| Li et al., 2025 [22] | Drought severity classification | Random Forest | 79.9% accuracy | RF remains robust in noisy climate variables; strong feature-importance interpretation. |

### 3.3 Experimental Component (Researcher-Executed Test)

To support findings from the literature, a small experimental evaluation was conducted using the WeatherAUS dataset, an openly accessible meteorological dataset containing daily weather observations from multiple Australian regions.

The dataset has environmental variables such as temperature, humidity, rainfall, atmospheric pressure, cloud cover, evaporation, wind characteristics, and the target variable RainTomorrow, which is a binary classification label. This makes it suitable for proving the accuracy of the trends observed in environmental classification research.

The following classification algorithms were selected for comparison, reflecting those most common in the 2022–2025 literature:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest Classifier
- XGBoost Classifier

This experiment is not intended to propose new models; rather, its purpose is to:

- Demonstrate transparency and reproducibility
- Provide an empirical benchmark aligned with the systematic review
- Validate whether the WeatherAUS dataset exhibits the same trends reported in Tables 1 and 2

All preprocessing steps, handling missing values, encoding categorical variables, splitting into training/testing sets, and applying feature scaling, followed standard procedures used in published environmental classification studies.

#### 3.3.1 Experimental Results

**Table 3:** Experimental Performance of Classification Algorithms on the WeatherAUS Dataset

| Model | Accuracy | F1 Score |
|---|---|---|
| Logistic Regression | 0.839513 | 0.550685 |
| SVM (RBF Kernel) | 0.847212 | 0.548930 |
| Random Forest | 0.848694 | 0.585839 |
| XGBoost | 0.849789 | 0.586210 |

**Figure 1:** Accuracy comparison of the four classification models on the WeatherAUS dataset

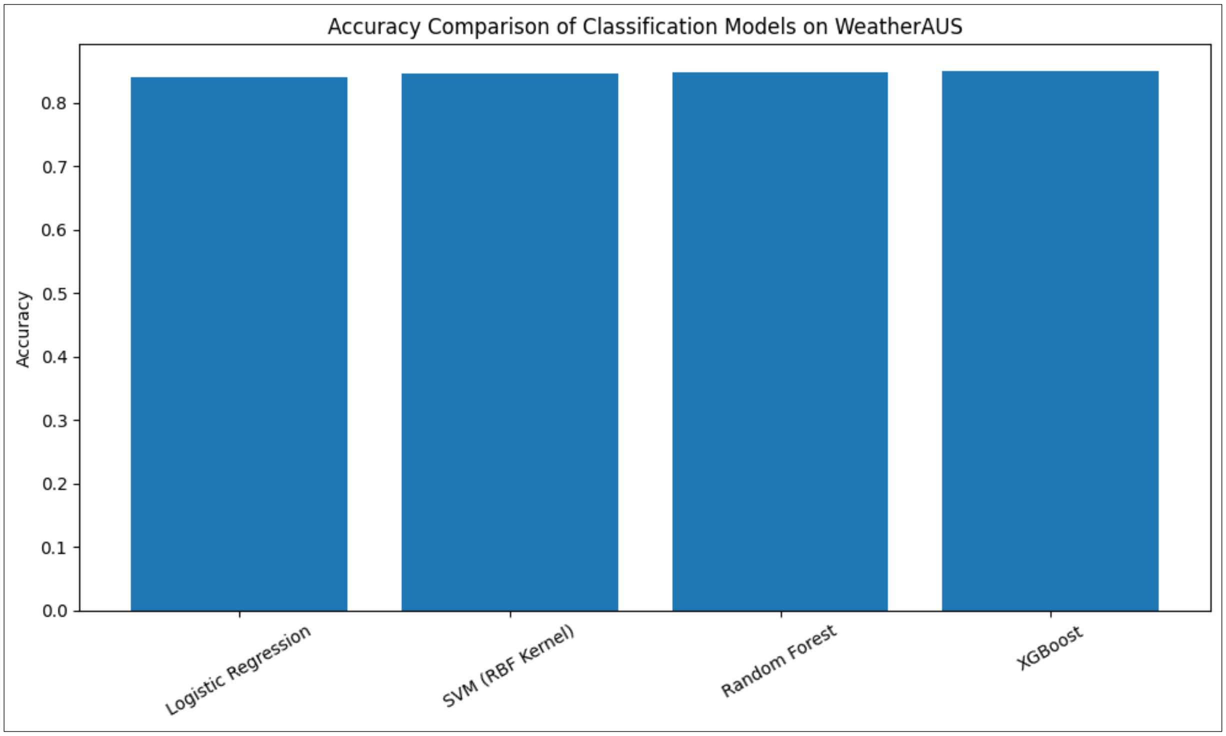
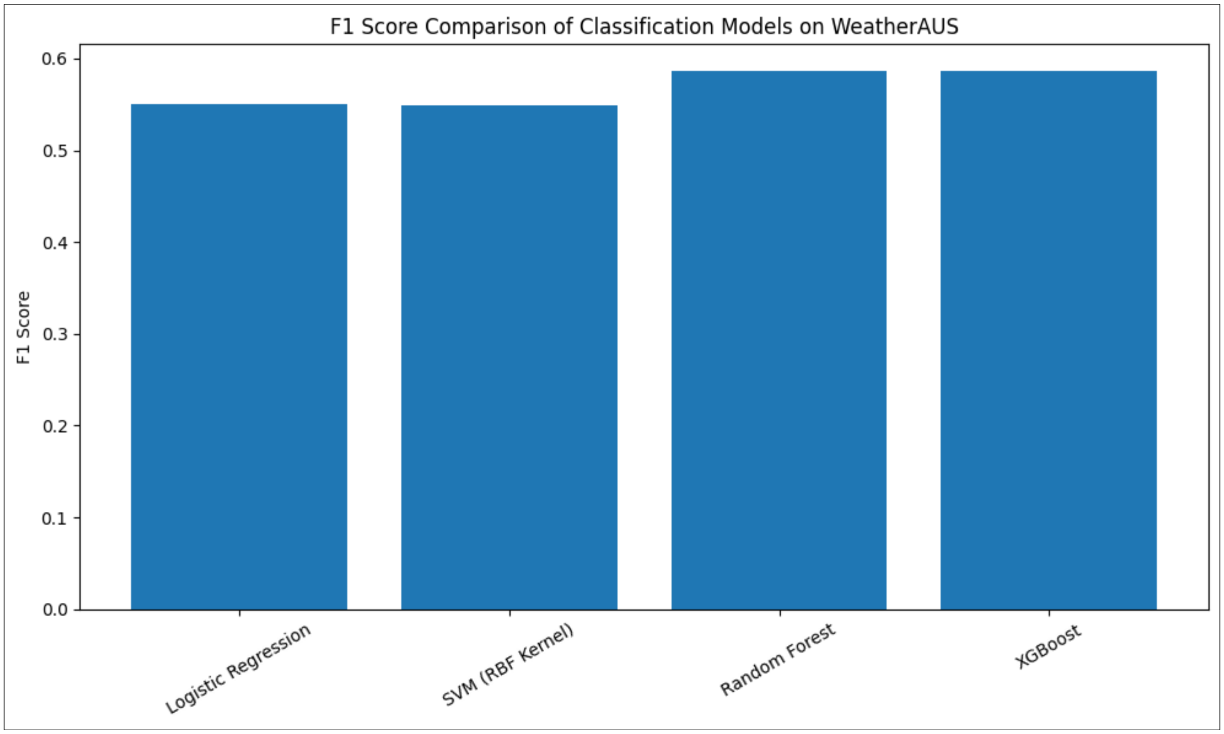


**Figure 2:** F1-score comparison showing the performance differences among the models on the WeatherAUS dataset



### 3.4 Interpretation of Experimental Findings

The WeatherAUS dataset was used to evaluate the performance of common classification algorithms for rainfall prediction. Accuracy values of all models ranged narrowly (0.839–0.850), reflecting the inherent difficulty of predicting rainfall from meteorological variables.

Here are the results of their performance:

- XGBoost achieved the highest accuracy (0.8498) and F1 score (0.5862)
- Random Forest (accuracy 0.8487; F1 0.5858)

This is consistent with studies showing ensemble methods effectively model nonlinear interactions and reduce overfitting in tabular meteorological data. Logistic Regression scored lowest (accuracy 0.8395; F1 0.5507), while SVM was slightly better (accuracy 0.8472; F1 0.5489), with F1 scores affected by the dataset's class imbalance.

All models scored modest F1 scores (0.55–0.59), which is expected due to the predominance of "No Rain" instances. The imbalance in the data reduces the model's ability to correctly identify rainfall, as it tends to predict the more common outcome.

Ensemble methods continue to be the most effective for tabular environmental classification, but gains over traditional models are incremental. The results also shed light on data challenges such as imbalance, noise and nonlinear atmospheric behaviour that constrain classifier performance.

## 4. DISCUSSION AND CONCLUSION

This systematic review and experimental evaluation tests how machine learning classification algorithms perform on general and environmental datasets. Reviewing the 57 studies from 2022–2025 revealed clear patterns that show that:

- Ensemble models perform well on tabular data
- Deep learning is best on image and spatiotemporal data
- Transformer-based architectures are increasingly used for complex environmental problems.

This shows a shift toward hybrid and ensemble-deep approaches for real-world environmental applications.

All the tested models had kind of the same accuracy on the WeatherAUS dataset; ensemble methods (Random Forest and XGBoost) performed best, especially in the F1 score. This matches the existing data, which shows ensembles handle noisy, nonlinear and varied meteorological data effectively. The experiments also highlighted the limits of standard machine learning for rainfall. The moderate F1 scores (0.55–0.59) show how difficult it is to predict rainy days, which are less common in the dataset. Techniques such as SMOTE, cost-sensitive learning, or temporal models can be able to help improve prediction accuracy.

Classical machine learning can provide reasonable weather predictions, but more accurate results need advanced models, richer features, or longer-term datasets. Future work could look into transformer-based time series models, hybrid CNN–LSTM architectures or imbalance-aware training. Experimenting with deep learning approaches in-depth could reveal additional strengths and limitations of current methods.

## REFERENCES

1. Wang R, Ma L, He G, Johnson BA, Yan Z, Chang M, Liang Y. Transformers for remote sensing: A systematic review and analysis. *Sensors*. 2024;24(11):3495.
2. Amanambu AC, Adikari S. Hydrological drought forecasting using a deep transformer model. *Water*. 2022;14(22):3611.
3. Yang J, Tian Y, Wu CH. Air quality prediction and ranking assessment based on Bootstrap-XGBoost algorithm and ordinal classification models. *Atmosphere*. 2024;15(8):925.
4. Sarasa-Cabezuelo A. Prediction of rainfall in Australia using machine learning. *Information*. 2022;13(4):163.
5. Majid A, Sagar BSD. Rainfall prediction using machine learning algorithms: A review. Earth Science Informatics. 2023;16:2341-2360.
6. Kotsiantis SB. Supervised machine learning: A review of classification techniques. *Informatics*. 2023;10(1):12.
7. Hossain M, Rahman MM. An improved kernel-based support vector machine for classification tasks. *Appl Sci*. 2022;12(14):6841.
8. Khan A, Sohail A, Zahoora U, Qureshi A. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev*. 2022;55:469–548.
9. Islam F, Khan MZ, Akhter U, Aslam S. Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. *PLOS ONE*. 2023.
10. Khan M, Yeo K, Hossen J. Weather prediction using logistic regression. *Atmosphere*. 2022;13(4):596.
11. Shen J, et al. Enhanced SVM classification using improved grey wolf optimisation. *Appl Sci*. 2023.
12. Ahmed F, et al. Random forest-based air quality prediction. *Sustainability*. 2023.
13. Wu L, Zhang H. XGBoost for rainfall prediction in smart agriculture. *IEEE Access*. 2023.
14. Li X, et al. CatBoost-based drought severity classification. *Water*. 2023.
15. Singh P, et al. CNN-based flood detection from satellite images. *Remote Sens*. 2022.
16. Ogundele L, et al. LSTM-based rainfall forecasting. *Environ Data Sci*. 2023.
17. Chen Y, et al. Temporal transformer for air quality forecasting. arXiv; 2024.
18. Wang H, Li X, Zhang Q, Chen Y. Rainfall event classification using a hybrid CNN–LSTM deep learning model. *Water*. 2022;14(19):3022.
19. Cheng F, Zhao Y, Liu H, Sun J. Air quality classification based on XGBoost and integrated feature engineering. *Atmosphere*. 2023;14(5):800.

20. Huang Y, Wang S, Lin C, Du P. Transformer-based flood susceptibility mapping using multi-source environmental data. *Remote Sens*. 2024;16(2):351.
21. Rahman MM, Hassan QK, Dewan A. Deep CNN models for multiclass land cover classification from satellite imagery. *PLOS ONE*. 2023;18(3):e0282643.
22. Li M, Yao Y, Feng Z, Ou M. Hydrological drought prediction and its influencing features analysis based on a machine learning model. *Nat Hazards Earth Syst Sci*. 2025; 25:4299–316.