Hybrid Anomaly-Based Intrusion Detection System for Zero-Day Exploits Using Machine Learning

Satvik Agrawal
Department of Information and Communication Technology
Manipal Institute of Technology, Manipal
Manipal, India
satvik2.mitmpl2022@learner.manipal.edu

Aprameya Ansh
Department of Information and Communication Technology
Manipal Institute of Technology, Manipal
Manipal, India
aprameya.mitmpl2022@learner.manipal.edu

***Abstract*—Zero-day attacks pose a significant threat to modern computer systems due to their unpredictable and previously seen nature. This study presents a hybrid model for anomaly detection to be implemented as part of an Intrusion Detection System (IDS) that combines a symmetric autoencoder and an XGBoost classifier for detecting previously unseen threats in real-time network environments. The approach involves preprocessing the UGRansome dataset, applying data cleaning and transformation methods such as skewness normalisation and categorical feature encoding. The proposed hybrid model offers an overall accuracy of 0.99, showcasing the capability of an autoencoder to act as a binary classifier for anomaly detection. The paired action of a reconstruction-based method and a supervised classifier for cross-verification improves the robustness of IDS solutions against zero-day threats.**

***Index Terms*—zero-day attack, anomaly detection, intrusion detection system (IDS), autoencoder, XGBoost, hybrid model, network intrusion detection system (NIDS), machine learning, unsupervised learning, cybersecurity**

I. Introduction

A zero-day attack exploits zero-day vulnerabilities. These weaknesses in a computer system are unknown to hosts (targets), and no immediate mitigation plan is geared toward them. Until these vulnerabilities are remedied, malicious actors could exploit such systems. Signature-based Intrusion Detection Systems (IDS) employ pattern matching to identify a known attack. This technique fails when no matching signature exists for an incoming threat [1]. Anomaly-based detection attempts to overcome the limitations of signature-based IDS by learning a computer system's 'normal behaviour' and identifying deviations as anomalies. Anomaly-based IDS solutions (AIDS) are classified as statistical, knowledge-based, and Machine-Learning (ML) based [2]. The advantage of this solution over signature-based IDS is its potential to identify unknown threats. The false positive rates of these anomaly-based techniques are generally high [3].

ML-based anomaly detection can be categorised into three broad classes [4]: *supervised*, *semi-supervised*, and *unsupervised*. In supervised learning, the model is trained on both labelled normal and anomalous data. Semi-supervised learning trains the model only on labelled normal data. Unsupervised learning does not require labelled training sets and operates under the assumption that anomalies are sparse. The core issue of supervised learning is that zero-day threats will not have a behavioural pattern similar to the anomalous data points in the training data. In real-world applications, anomalies are relatively not sparse and exhibit behaviour similar to benign data. The lack of a generalised solution against zero-day threats can be attributed to the development of overfitted and biased predictive classifiers, and a lack of public datasets composed of recent attack patterns [5].

The Global-Malware Volume Statistics [6] (recorded from 2015 to March 2020) showcased the rapidly evolving nature of unknown threats. It is assumed that an ML model trained on 5 years of data would be capable of mitigating the attacks known to that date. After the model was trained till March 2020, there were 240.52 million new attacks that the model was unaware of. The model was not able to classify them due to a lack of labelling, and as a result, the model's accuracy dropped significantly [5]. Unsupervised anomaly detection methods cannot compensate for the volatile nature of zero-day threats. Paired with the lack of labelled anomalous data for training, even Deep Learning (DL) based AIDS fail to identify unknown threats in real-world applications.

A hybrid model combines supervised and unsupervised learning to detect anomalies without solely relying on labelled training data. It overcomes the potential risks of a high False Positive Rate (FPR) through cross-verification with a supervised classifier [7]. This study proposes a hybrid model for anomaly detection, pairing an autoencoder with an XGBoost classifier for robust zero-day threat identification. The autoencoder is implemented as a binary classifier, wherein it is trained to reconstruct normal data accurately and has high reconstruction errors for predicting unknown data. The comparison of these reconstruction errors to a threshold value serves as the basis for classification.

The remainder of the paper is composed as follows. We aim to substantiate our efforts through a comprehensive literature review (Section II) of articles published in the past five to ten years. Despite the extensive research publicly available, the success of such IDS solutions in real-time operational environments is very limited [8]. We aim to mitigate this knowledge gap through the review of application-oriented

studies with an interest in ML and DL technique integration, prioritising unsupervised learning models. Section III details the architecture and process flow of our proposed model, preluded by a summarisation of the dataset employed for model training. Section IV provides result metrics and evaluates the codebase for project implementation. Section VII presents future trends and challenges to countering zero-day threats in real-time applications. Section VI concludes this study.

II. Review of Existing Work

A literature survey was conducted to analyse pre-existing hybrid models for anomaly detection. Peer-reviewed papers published in the past five years were investigated, and the search was refined using appropriate keywords. Publications were reviewed, keeping in mind the nature of the input feature set to determine the appropriate usage of an autoencoder for anomaly detection.

Sommer's *et al.* [8] study on using ML for Network Intrusion Detection focused on the significance of ML-based techniques for intrusion detection. The study highlighted the high cost of errors (false positives, false negatives) in intrusion detection. In the study, a set of guidelines was formulated with the goal of strengthening future research in anomaly detection. The study emphasised the role of domain knowledge in avoiding the duplication of efforts in IDS solutions. We infer that when developing such a system, the chosen classifier or predictive model cannot be independent of the context in which it is being applied; Performance metrics cannot solely be relied on for a generalised solution. In their survey of IDS solutions, Khraisat *et al.* [2] reviewed network traffic datasets for intrusion detection (public datasets such as DARPA, KDD, NSL-KDD). A summary of these datasets was provided, including their viability in modern-day IDS solutions, their limitations, and their contributions to IDS. These datasets were compared concerning (a) the availability of labelled data, (b) the inclusion of zero-day attacks, and (c) their significance in ML-based anomaly-detection techniques.

In Kaur and Singh's [1] proposed hybrid zero-day attack detection model, a 3-layer architecture was detailed, composed of a combination of signature-based and anomaly-detection systems. The anomaly-detection method constituted a 1-class Support Vector Machine (SVM), intended to be able to identify test data that is not alike the training data. The motivation behind a 1-class SVM was to classify the data as either 'known' (known signature) or 'unknown' (anomalous). Yi-Xuan Xu's *et al.* [9] reconstruction-based detection model with a RandomForest classifier showcased the performance of a hybrid anomaly-detection model, and subsequently highlighted the limitations of the isolated implementation of supervised and unsupervised models, respectively. This 'RecForest' model presented its novelty and contribution as having minimal training costs and hyperparameter tuning requirements. Torabi's *et al.* [10] proposed autoencoder model utilising vector reconstruction error presents the innovative use of a vector of reconstruction errors in place of a singular value derived from the trained autoencoder. The model trained a set of autoencoders, each on a distinct class of data (Normal, Attacker, Unknown, etc.), utilising each autoencoder as a binary classifier for the respective classes. The study highlighted the usage of reconstruction losses derived from autoencoder-generated samples as an anomaly detector.

A. Insights and Knowledge Gaps

The works surveyed above highlight the importance of context-aware modelling and the significance of the dataset employed. However, the studies do not specify the appropriate means of handling categorical data and instead provide examples. This may be due to the lack of successfully applied IDS solutions. Our proposed model architecture integrates the above guidelines and improves upon feature engineering methods as well as efficient dataset utilisation through the use of a robust dataset (detailed in Section III).

The use of an autoencoder as a binary classifier for anomaly detection is validated by the exemplary results showcased by previous works [10]. This implementation of an autoencoder is computationally inexpensive when compared to a segmented implementation of One-Class Support Vector Machines (OCSVM) [11] and highly-tuned RandomForest classifiers [9].

III. Proposed Methodology

A. Dataset Description and Preprocessing

Both the autoencoder and supervised classifier were trained on the UGRansome dataset. The UGRansome dataset [12] has 14 attributes, of which 6 are numerical and 8 are categorical (including the target variable), as briefly described in Table I. The 'Time' column is composed of timestamps and includes negative values, which were linearly shifted to avoid misinterpretation of temporal data. The categorical attributes 'SeddAddress', 'ExpAddress', and 'Port' were dropped due to their irrelevance when identifying anomalies. 'BTC', 'USD', and 'Netflow Bytes' are numerical attributes exhibiting high skewness and were non-linearly transformed to reduce effect of abnormality on model performance. The target variable 'Prediction' consists of 3 unique class labels: Anomaly (A) data for training predictive models against zero-day threats, Signature (S) data for evaluating signature-based IDS solutions and providing a reference for training classifiers, and Synthetic Signature (SS) data provided for training supervised anomaly and behaviour-based detection models. Non-linear transformations were applied on numerical attributes ('USD', 'BTC', 'Netflow Bytes') to counter the skewness, under the assumption that the attributes were unimodal in nature (verified by histogram analysis). The guidelines and rules for skewness analysis and normalisation were derived from the publication by Max Kuhn *et al.* [13] titled "Applied Predictive Modeling".

An appropriate feature encoding method was employed referencing publicly available guidelines [14] for categorical variable encoding. The cardinality of the categorical variables was calculated, based on which the appropriate encoding

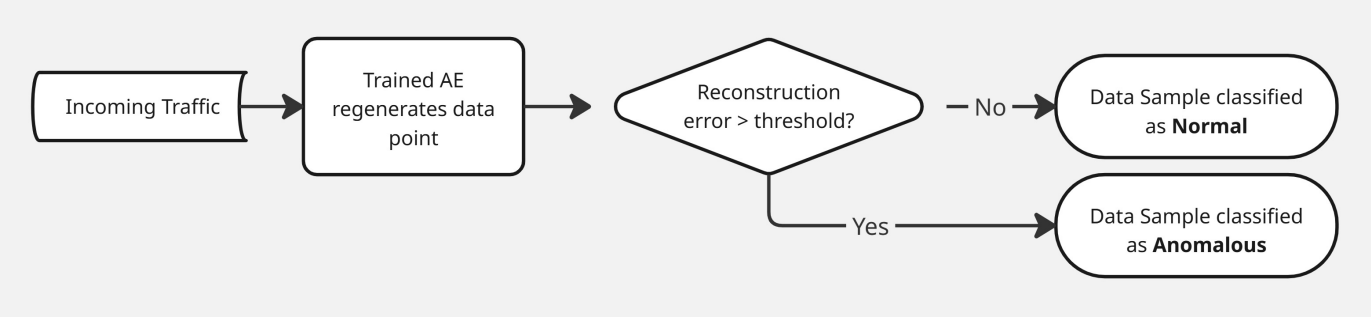

Fig. 1. Implementation Flow of Autoencoder

scheme was employed. The chosen method was *One-Hot Encoding* considering the relatively low cardinality (<15) of categorical features, and the compatibility of one-hot encoding with reconstruction-based methods. Min-max scaling was applied on the encoded feature set to ensure compatibility with the autoencoders activation function [15] and avoid poor reconstruction loss generation from the Binary Cross-Entropy loss function [16].

B. Model Architecture

The autoencoder is symmetric and compresses the input data into an 8-dimensional feature set through a set of two Dense layers ($128 \rightarrow 64$), each undergoing L2 Regularization, Batch Normalization, and LeakyReLU activations. It reconstructs the data through mirrored Dense layers and then passes them through sigmoid-activated output layer. This outputs a set of reconstruction errors in the range of $[0, 1]$. The reconstruction loss function is Binary-Cross Entropy.

The autoencoder is trained on normal data composed of Signature (S) and Synthetic Signature (SS) unlabelled data. The model learns to reconstruct normal data accurately and have relatively high reconstruction losses for anomalous data. The autoencoder reconstructs incoming traffic. If the reconstruction error is greater than the threshold, the data is classified as anomalous (showcased in Figure 1).

TABLE I
DATASET ATTRIBUTE DESCRIPTIONS

Attribute	Description
Time	Timestamps of network attacks
Protocol	Network protocol used
Flag	Network connection status
Family	Type of network intrusion
Clusters	Groupings of events
SeddAddress	Ransomware attack URLs
ExpAddress	Original attack URLs
BTC	Bitcoin transactions in attacks
USD	Financial damages (USD)
Netflow Bytes	Network flow byte size
IPaddress	IPs involved in events
Threats	Nature of threats
Port	Port number used
Prediction	Target label: anomaly (A), signature (S), synthetic signature (SS)

TABLE II
F1-SCORE PER CLASS FOR EACH CLASSIFIER

Class	XGBoost	Logistic Regression	KNN
A	0.99	0.83	0.98
S	0.99	0.89	0.98
SS	0.99	0.89	0.99

The threshold is identified by evaluating the Receiver Operating Characteristic (ROC) curve, used to illustrate the performance of a binary classifier at varying threshold values [17]. An alternate method was evaluated utilising Youden's J Statistic (also called Youden's Index) [18]. Ultimately, the threshold identified as having a 99% recall rate for anomalous data was evaluated to be the most efficient.

For the supervised classifier models, a comparison was drawn between an XGBoost Classifier, a LogisticRegression model, and a K-Nearest Neighbours (KNN) classifier with respect to their F1-scores for the target class labels. The results of this evaluation can be seen in Table II. All models were evaluated on an 80-20 train-test split on the entire labelled dataset. The XGBoost Classifier showed an exemplary F1-score of 0.99 for all class labels, outperforming the alternatives. The detailed confusion matrices and result metrics are available in the Appendix A.

IV. RESULT METRICS AND ANALYSIS

A. Metrics for Evaluation

Some evaluation metrics are used multiple times in this Section and are defined as such:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where *TP* refers to True Positives, *FP* refers to False Positives, and *FN* refers to False Negatives in the prediction. High Precision implies that the model rarely wrongly classifies something as positive. A high Recall value suggests that the model detects most of the actual positive cases. The F1-score is the harmonic mean between Precision and Recall.

The performance of the data transformation techniques was quantified by the skewness value before and after iterative normalisation.

The autoencoder's performance was evaluated by making it reconstruct a test-set of normal and anomalous data. They were classified as normal or anomalous based on their reconstruction errors (as described in Figure 1). For comparison, we use the mapping: Normal $\rightarrow$ 0, Anomaly $\rightarrow$ 1. This mapping was applied on the reconstructed data samples, which are the predictions of the autoencoder. These predictions were compared to the true labelled values of the test-data (known to us) for evaluating the model performance. The performance of the XGBoost classifier was also evaluated based on the aforementioned metrics.

For the autoencoder, the distinction between the distribution of reconstruction errors for normal and anomalous data was reviewed through empirical metrics such as the Wasserstein Distance [19] and Kolmogorov-Smirnov (KS) statistic [20].

B. Results

TABLE III
RESULTS OF SKEWNESS NORMALIZATION

Attribute	Before Normalization	After Normalization
USD	3.232	-1.918
BTC	11.936	0.086
Netflow Bytes	1.573	-0.113

TABLE IV
PERFORMANCE OF AUTOENCODER

Class	Precision	Recall	F1-Score
Normal (S/SS)	0.98	0.99	0.99
Anomaly (A)	0.99	0.99	0.99

TABLE V
PERFORMANCE OF XGBOOST CLASSIFIER

Class	Precision	Recall	F1-Score
A	0.99	0.99	0.99
S	0.99	0.99	0.99
SS	0.99	0.99	0.99

The Wasserstein distance evaluates the difference between two probability distributions, but can be misinterpreted to solely mean separation between distributions. A value of **0.0279** indicates that the two distributions may share characteristics, but not necessarily that they overlap. Paired with this, a high KS-statistic of **0.7072** (and a p-value of 0.000) implies that the distributions diverge significantly at some point. As visualised in Figure 2, there is a significant overlap until a certain point, after which the distributions diverge.

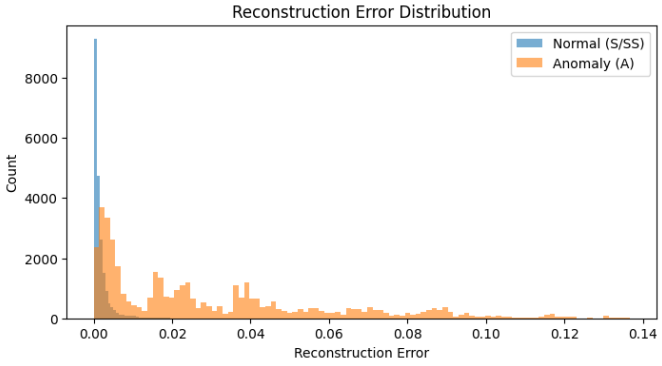

Fig. 2. Distribution of Reconstruction Errors for Normal and Anomalous Data

V. DISCUSSION

The result of the data transformation techniques applied as detailed in Table III indicates the effectiveness of reducing skewness and avoiding the presence of abnormal data in the predictive modelling training set.

The performance of the Autoencoder (Table IV) was quantified by the precision, recall, and F1-scores for the binary classes. The precision of 99% for anomalous data indicated that the model was correctly identifying positively-labelled anomalies, and the recall value implied that 99% of true anomalies were actually getting detected. These results showcase the exemplary capability of the autoencoder model in identifying unknown threats.

The accuracy of the XGBoost model as a multi-class classifier (Table V) was validated by the near-perfect accuracy of 99%. The potential of overfitting was countered by evaluating the F1-score for all classes as well, which also showcased 0.99. The number of *FP*s and *FN*s were evaluated as well (Figure 4 in Appendix A), indicating that the accuracy score was trustworthy.

VI. CONCLUSION

The study validates the effectiveness of data preprocessing techniques concerning the contextual utilisation of feature sets. The novel algorithm developed for iteratively reducing skewness in unimodal numerical attributes proved effective in normalising the data.

The study validates the performance of a hybrid model in place of isolated supervised or unsupervised methods for anomaly detection. The capability of an autoencoder as a binary classifier was showcased through an exemplary model accuracy of 99%.

VII. FUTURE SCOPE

Due to the limited number of features in the training dataset (14 total, of which 3 were vestigial), feature selection methods

were not employed or evaluated. The same may not be true for other datasets (such as KDD). Evaluation of the hybrid model's performance may be done with a high-dimensionality dataset. The requirement of feature selection may be evaluated through model performance as well. The integration of a preluding Signature-based detection layer (for example, one that employs SNORT rules) may be done to enhance overall performance of the IDS.

REFERENCES

[1] R. Kaur and M. Singh, "A hybrid real-time zero-day attack detection and analysis system," *International Journal of Computer Network and Information Security*, vol. 7, no. 9, pp. 19–31, 2015.

[2] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, p. 20, Jul 2019.

[3] A. Alazab, M. Hobbs, J. Abawajy, and M. Alazab, "Using feature selection for intrusion detection system," in *2012 International Symposium on Communications and Information Technologies (ISCIT)*, pp. 296–301, 2012.

[4] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," *IEEE Access*, vol. 9, pp. 78658–78700, 2021.

[5] R. Ahmad, I. Alsmadi, W. Alhamdani, and L. Tawalbeh, "Zero-day attack detection: a systematic literature review," *Artificial Intelligence Review*, vol. 56, pp. 10733–10811, Oct 2023.

[6] Statista, "Cumulative detections of newly-developed malware applications worldwide from 2015 to march 2020." https://www.statista.com/statistics/680953/global-malware-volume/, Aug 2020.

[7] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, p. 41, Jul 2020.

[8] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," *IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.

[9] Y.-X. Xu, M. Pang, J. Feng, K. M. Ting, Y. Jiang, and Z.-H. Zhou, "Reconstruction-based anomaly detection with completely random forest," in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 127–135, SIAM, 2021.

[10] H. Torabi, S. L. Mirtaheri, and S. Greco, "Practical autoencoder based anomaly detection by using vector reconstruction error," *Cybersecurity*, vol. 6, no. 1, p. 1, 2023.

[11] G. Pu, L. Wang, J. Shen, and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," *Tsinghua Science and Technology*, vol. 26, no. 2, pp. 146–153, 2021.

[12] M. N. W. Nkongolo, "Ugransome dataset," 2023.

[13] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*, vol. 26. Springer, 2013.

[14] Datafreak, "All about categorical variable encoding, towards data science." https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02, 2023. Accessed: 2025-04-15.

[15] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[17] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[18] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[19] V. M. Panaretos and Y. Zemel, "Statistical aspects of wasserstein distances," *Annual review of statistics and its application*, vol. 6, no. 1, pp. 405–431, 2019.

[20] M. Andriulli, J. K. Starling, and B. Schwartz, "Distributional discrimination using kolmogorov-smirnov statistics and kullback-leibler divergence for gamma, log-normal, and weibull distributions," in *2022 Winter Simulation Conference (WSC)*, pp. 2341–2352, IEEE, 2022.

APPENDIX A
RESULT GRAPHICS AND CHARTS

TABLE VI
CONFUSION MATRIX

Autoencoder	
TP = 42135	FN = 426
FP = 213	TN = 21084

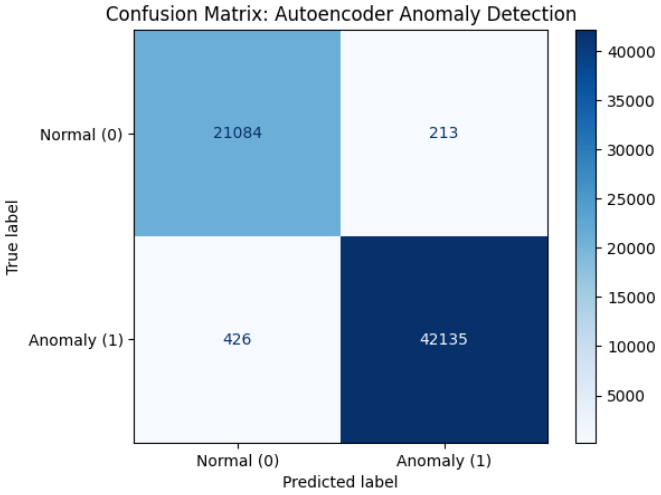

Fig. 3. Confusion Matrix Heatmap for Autoencoder

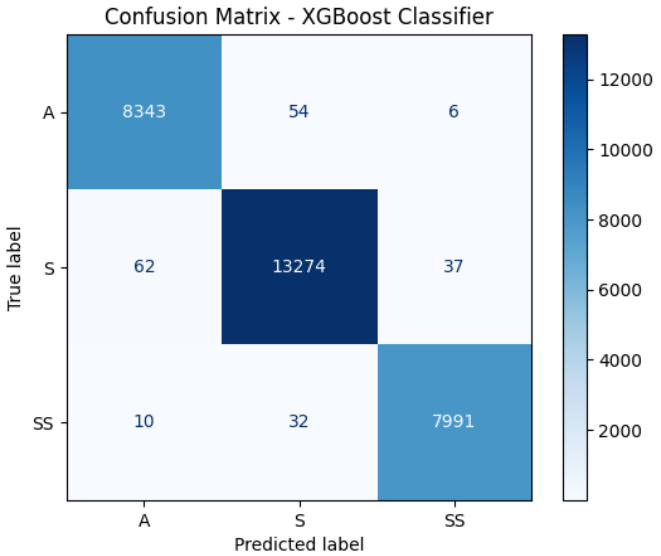

Fig. 4. Confusion Matrix Heatmap for XGBoost Classifier

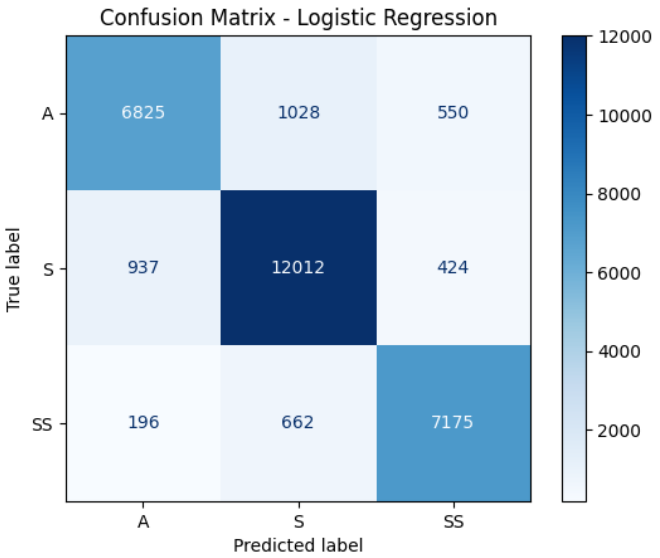

Fig. 5. Confusion Matrix Heatmap for Logistic Regression Classifier

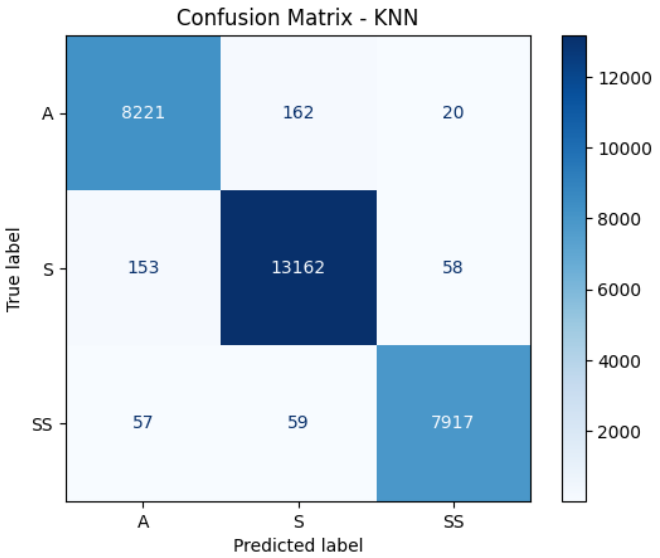

Fig. 6. Confusion Matrix Heatmap for K-Nearest Neighbours Classifier