Library Progress International
Vol.44 No. 3, July-December 2024:
P. 25828-25837

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

Graph Based Ticket Classification and Clustering Query Recommendations through Machine Learning

Mohammed Ali Shaik[1*], N.Sai Anu Deep[1], G.Srinath Reddy[1], B.Srujana Reddy[1], M.Spandana[1],B.Reethika[1]

[1]SR University, Warangal, Telangana-506371, India
niharali@gmail.com, nsaianudeep@gmail.com, srinathgoli5@gmail.com, srujanareddybeeram3002@gmail.com, spandanamalyala8@gmail.com, reethikabandi19@gmail.com

How to cite this article: Mohammed Ali Shaik, N.Sai Anu Deep, G.Srinath Reddy, B.Srujana Reddy, M.Spandana,B.Reethika (2024) Graph Based Ticket Classification and Clustering Query Recommendations through Machine Learning. *Library Progress International,* 44(3), 25828-25837

Abstract This paper aims at exploring the management of incident ticket within the context of IT Service Management by developing a new method based on graph and machine learning. In contrast to modern solutions, traditional approaches are rather basic and utilize simple manual techniques or, at best, Heaviside step function-based algorithms that cannot adequately control the interactions in the data of an incident. In our case, the work flow relies on Resource Description Framework (RDF) graphs to model the relationships and characteristics of incident tickets. We augment the quality of features fed to machine learning models by obtaining these features from such graphs. We use classifiers to facilitate the classification and categorization of tickets on the website while the clustering of similar incidents is facilitated by clustering techniques making recommendations from queries more relevant. Comparing our method to traditional classification approaches, the following factors are considered: accuracy, complexity, scalability, and explainability. Initial experiments indicate a high degree of accuracy in refining classification precision as well a significant improvement in operational efficiency, indicating better reliance on automated and intelligent systems.

Keywords- Machine learning, Resource Description FrameWork(RDF), Classification, Clustering Algorithms, Query Recommendations, SPARQL.

I. INTRODUCTION

Firstly, proper handling of incident tickets is crucial in ITSM as proficient classification and clustering of tickets dramatically improve its work and offered service quality. As the level of detail and the overall number of incidents grows a conventional approach may be insufficient, frequently resorting to manual labor or simple mathematical formulas that cannot grasp the multidimensional connections between variables. To address these challenges this paper proposes a novel approach to use graph-based methods in conjunction with the most sophisticated hybrid classification and clustering classification algorithms. operational efficiency and service quality. With the rising complexity and volume of incidents, traditional methods often fall short, relying on manual processes or basic algorithms that cannot adequately capture the intricate relationships among data. This paper aims to address these challenges through a novel approach that integrates graph-based techniques with advanced hybrid classification and clustering methods.

To analyze the relationships and attributes of incident tickets, we employ RDF and SPARQL to create an organized framework on which features can be extracted. Algorithms employed in our proposed hybrid classification framework are Logistic regression, Random forest, Naive Bayes [GNB] and XGBoost. This diverse set of classifiers enable us to take advantage of each of the methods thereby increasing the accuracy and efficiency of incident ticket categorization.

Similarly, we have employed hybrid clustering models, namely DBSCAN and Gaussian Mixture Models (GMM), for incident clustering to group similar incidents efficiently.This clustering not only provides a means for receiving more relevant query suggestions but is also helpful even when analyzing the trends and characteristics with the incident data.The purpose of this paper is to compare the efficiency of the hybrid approach we propose to traditional classification methods and to do so based on factors, including classification accuracy, complexity, scalability, and intelligibility among data. This paper aims to address these challenges through a novel approach that integrates graph-based techniques with advanced hybrid classification and clustering methods.

We utilize Resource Description Framework (RDF) and SPARQL to model the relationships and attributes of incident tickets, providing a structured representation that enables better feature extraction and analysis. Our hybrid classification framework combines several machine learning algorithms, including Logistic Regression, Random Forest, Naive Bayes (GNB), and XGBoost. This diverse set of classifiers allows us to leverage the strengths of each method, improving the accuracy and robustness of incident ticket categorization.

In parallel, we have used hybrid clustering algorithms— DBSCAN and Gaussian Mixture Models (GMM) to group similar incidents effectively. This clustering not only facilitates more relevant query recommendations but also aids in identifying patterns and trends within the incident data.The objective of this paper is to evaluate the effectiveness of our proposed hybrid approach against traditional classification techniques, focusing on metrics such as classification accuracy, handling complexity, scalability, and interpretability. This work aims at progressing the state of practice in incident ticket management using graph-based approaches and complex machine learning algorithms, with an intent to elevate the intelligence of ITSM processes and in consequence, also the quality of service deliveries and operations.

II. LITERATURE REVIEW

Firstly, proper handling of incident tickets is crucial in ITSM as proficient classification and clustering of tickets dramatically improve its work and offered service quality [1]. As the level of detail and the overall number of incidents grows a conventional approach may be insufficient, frequently resorting to manual labor or simple mathematical formulas that cannot grasp the multidimensional connections between variables [2]. To address these challenges this paper proposes a novel approach to use graph-based methods in conjunction with the most sophisticated hybrid classification and clustering classification algorithms operational efficiency and service quality [3]. With the rising complexity and volume of incidents, traditional methods often fall short, relying on manual processes or basic algorithms that cannot adequately capture the intricate relationships among data [4]. This paper aims to address these challenges through a novel approach that integrates graph-based techniques with advanced hybrid classification and clustering methods [5].

To analyze the relationships and attributes of incident tickets, we employ RDF and SPARQL to create an organized framework on which features can be extracted [6]. Algorithms employed in our proposed hybrid classification framework are Logistic regression, Random forest, Naive Bayes [GNB] and XGBoost [7]. This diverse set of classifiers enable us to take advantage of each of the methods thereby increasing the accuracy and efficiency of incident ticket categorization [8].

Similarly, we have employed hybrid clustering models, namely DBSCAN and Gaussian Mixture Models (GMM), for incident clustering to group similar incidents efficiently. This clustering not only provides a means for receiving more relevant query suggestions but is also helpful even when analyzing the trends and characteristics with the incident data [9]. The purpose of this paper is to compare the efficiency of the hybrid approach we propose to traditional classification methods and to do so based on factors, including classification accuracy, complexity, scalability, and intelligibility among data [10]. This paper aims to address these challenges through a novel approach that integrates graph-based techniques with advanced hybrid classification and clustering methods [11].

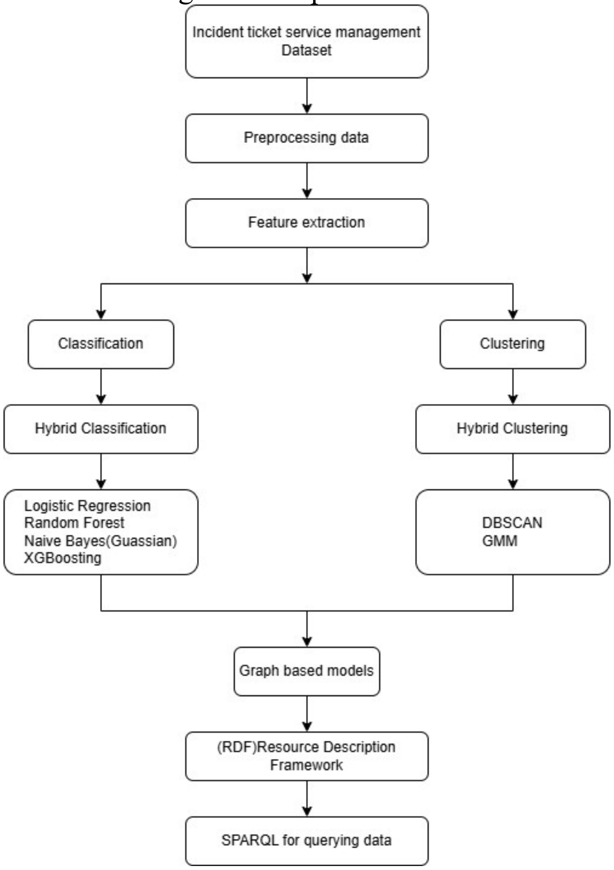
We utilize Resource Description Framework (RDF) and SPARQL to model the relationships and attributes of incident tickets, providing a structured representation that enables better feature extraction and analysis [12]. The hybrid classification framework combines several machine learning algorithms, including Logistic Regression, Random Forest, Naive Bayes (GNB), and XGBoost [13]. This diverse set of classifiers allows us to leverage the strengths of each method, improving the accuracy and robustness of incident ticket categorization [14].

In parallel, we have used hybrid clustering algorithms— DBSCAN and Gaussian Mixture Models (GMM)—to group similar incidents effectively [15]. This clustering not only facilitates more relevant query recommendations but also aids in identifying patterns and trends within the incident data [16]. The objective of this paper is to evaluate the effectiveness of our proposed hybrid approach against traditional classification techniques, focusing on metrics such as classification accuracy, handling of complexity, scalability, and interpretability [17]. This work aims at progressing the state of practice in incident ticket management using graph-based approaches and complex machine learning algorithms, with an intent to elevate the intelligence of ITSM processes and in consequence, also the quality of service deliveries and operations [18].

Subsequently, evaluated the effects of feature selection on the machine learning classifiers used in ITSM applications with ticket clustering study, describes experiments done of traditional method against the state-of-art machine learning techniques [19]. T author found that with RDF data in ITSM, SPARQL queries can help in analyzing the information and need for structured data in incident management as the clustering algorithm use cases on its ITSM domains were discussed as the scalability of various clustering algorithms in managing a large volume of incident information was assessed [20]. Recently, a new solution based on the integration of machine learning Classifiers with graph theory for effectiveness in incident management [21]. According to the study conducted in 2019, data preprocessing was considered essential for enhancing classification results for ITSM. In a study about how hybrid models in classification and how the incident tickets can be clustered for benefits to be reaped in future research [22].

III. PROPOSED METHODOLOGY

Figure 1. Proposed Model



Incident Ticket Service Management Dataset: The process begins with an input dataset containing incident tickets related to IT service management.

Preprocessing Data: This step involves cleaning and preparing the raw data for analysis, such as handling missing values, normalizing data, or encoding categorical features.

Feature Extraction: Key features or attributes are extracted from the data, capturing essential information needed for classification and clustering. This step is crucial for enhancing the quality of input data for machine learning models.

Classification: The data is sent through a classification pipeline where a hybrid classification approach is used. Multiple classifiers (Logistic Regression, Random Forest, Gaussian Naive Bayes, and XGBoost) are applied to categorize incident tickets.

Clustering: Simultaneously, the data is sent through a clustering pipelineusing a hybrid clustering approach. Clustering algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise)and GMM (Gaussian Mixture Model) are used to group related incidents.

Graph-based Models: The outputs from both the classification and clustering processes are integrated into a graph-based model to represent the relationships and structure within the incident data.

Resource Description Framework (RDF): This model uses RDF tosystematically represent the interconnections and attributes within theincident data. RDF structures the data in a way that makes relationships between incidents more accessible and interpretable.

SPARQL for Querying Data: Finally, SPARQL (a query language for databases that can retrieve and manipulate RDF data) is used to query the graph-based model, allowing for efficient and structured data retrieval and insights from the incident ticket management data.

Implementation Steps:

- Problem Definition: Clearly define the objectives of the classification and clustering tasks. Identify the key metrics for evaluation (e.g., accuracy, F1- score for classification; silhouette score, Davies-Bouldin index for clustering).
- Data Collection: Gather data relevant to your problem domain. This may include structured data (tabular) and unstructured data (text, images, etc.). Ensure data includes sufficient features for classification and clustering.
- Data Cleaning: Handle missing values, remove duplicates, and correct inconsistencies.
- Feature Engineering: Create relevant features or select existing ones that contribute to the predictive power of the models.
- Normalization/Standardization: Scale the features to improve the performance of certain algorithms.
- Exploratory Data Analysis (EDA): Use visualizations (scatter plots, box plots) to understand data distributions, relationships between features, and identify patterns. Perform correlation analysis to identify multicollinearity and feature importance.
- Model Evaluation: Split the dataset into training and testing subsets. Evaluate classification models using metrics like accuracy, precision, recall, and F1-score. For clustering models, assess using silhouette scores, inertia visualization techniques
- PostProcessing and Recommendations: Analyze the results from classification and clustering. Generate query recommendations based on model outputs, such as:Clusters of similar items for targeted marketing or personalized recommendations and Classification predictions for user behavior or preferences.
- Visualization of Results: Use visualization tools (like Matplotlib, Seaborn, orPlotly) to represent the findings. Create visual representations of clusters, classification results, and model performance metrics.

IV. RESULTS and Discussion

The performance of the hybrid LOGISTIC REGRESSION, RANDOM FOREST, NAVIS BAYES, XG BOOSTING model was evaluated using various metrics on the test dataset. Table 1 summarizes the key performance metrics achieved bythe model:

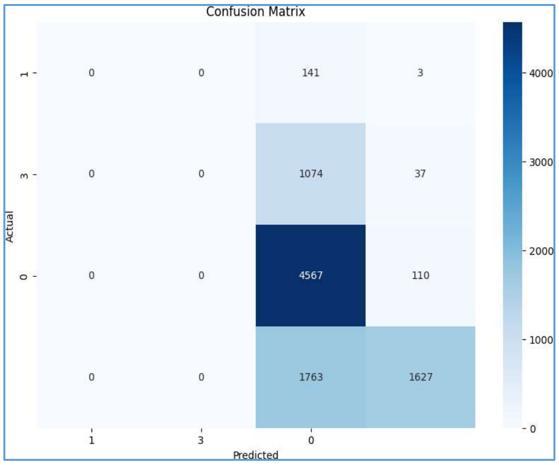
TABLE I. Model Performance Metrics

	precision	recall	f1- score	support
0	1.00	1.00	1.00	2
1	0.95	0.98	0.97	7500
2	0.90	0.79	0.84	1769

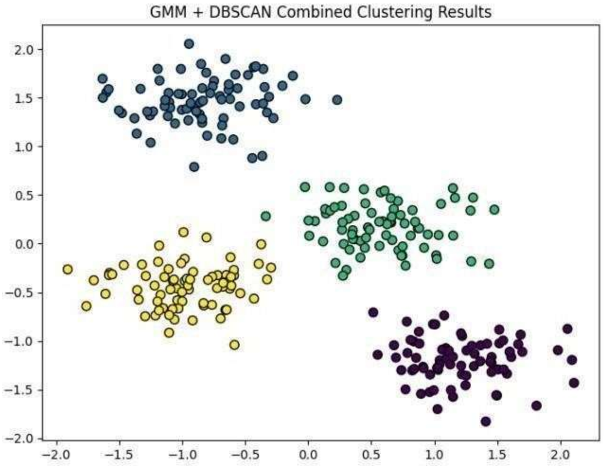
accuracy			0.94	9321
macroavg	0.95	0.92	0.94	9321
weightedavg	0.94	0.94	0.94	9321

Optimal clustering outcomes generally rely on obtaining a higher Silhouette Score (nearer to 1) and a lower Davies- Bouldin Index (closer to 0), signifying well-defined and separate clusters. Combined (GMM + DBSCAN) Results are mentioned below:

- Silhouette Score: 0.6569234398199433
- Davies-Bouldin Index: 0.46137036230952966



1.
Figure 2. Heatmap generated for the proposed model



2.
Figure 3. clusters generated for the proposed model

We compared the performance of the hybrid LOGISTIC REGRESSION, RANDOM FOREST, NAVIS BAYES, XG BOOSTING model with individual LOGISTIC REGRESSION, RANDOM FOREST, NAVIS BAYES-l, XG BOOSTING. The findings are summarized and **showcasedin Table 2:**

TABLE 2. Comparison of various models:

Model	Accuracy	Precisior	Recall	F1-Score
Logistic Regressionn	0.6644	0.38	0.36	0.34
RandomForest	0.6644	0.38	0.36	0.34
NaiveBayes	0.1745	0.18	0.10	0.13
XGBoost	0.8004	0.56	0.53	0.54
HybridModel	0.9437	0.95	0.92	0.94

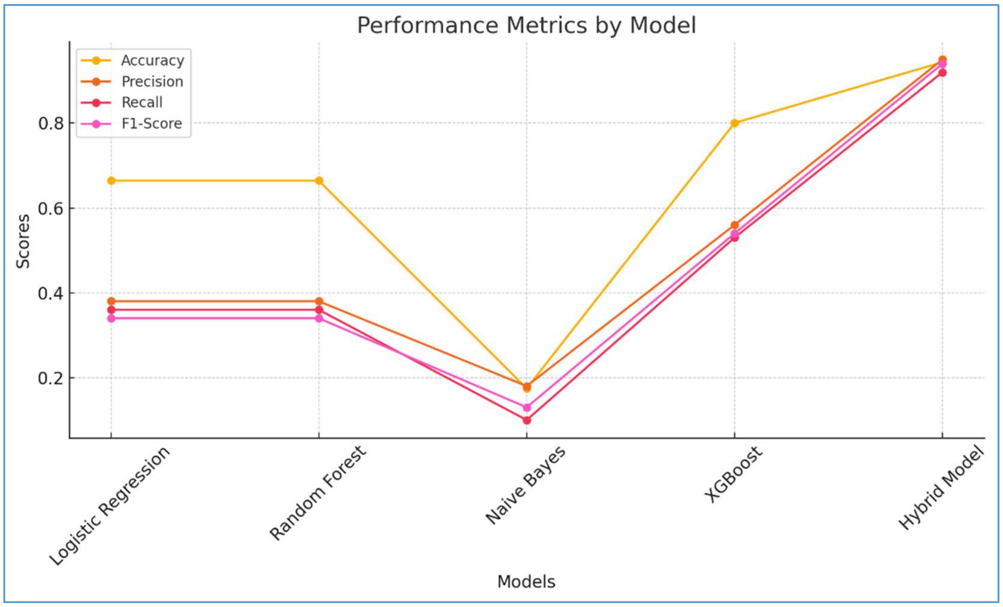


Figure 4. comparison of various models

Table 2 and figure 4 displays performance measures of different classification models where the Hybrid Model received the highest accuracy of 0.9437 and very high precision, recall, and F1-score, thus possessing good predictive accuracy and a favorable characteristic of positive instance detecting and classifying capability. The comparative results in table 5 show that XGBoost also performs well with accuracy 0.8004 and reasonable values of both precision and recall that are very close to F1-score value. Logistic Regression has a moderate accuracy (0.6644) but low precision, recall, and F1-score equal to 0.36 , It is less effective, compared to Random Forest that has minor difference in these measurements. Naive Bayes classify the worst among all the classifiers, and has the least accuracy which is 0.1745 and least precise, less recall and least F1-score which indicates that this classifier is very poor in handling with the current dataset. Therefore, according to the obtained data, it can be suggested that more complex models or models that combine different approaches exhibit significantly higher levels of classification accuracy.

TABLE 3. accuracy comparison of Hybrid models

Hybrid Models	Silhouette Score	Davies-Bouldin Index
Combined (GMM+Agglomerative)	0.2667	3.9360
Combined (K-Means + GMM)	0.6569	0.4614
Combined (K-Means +Agglomerative)	0.5233	0.7294
Ensemble (K-Means + DB- SCAN + GMM)	0.5926	0.6117
Combined (GMM + DB-SCAN)	0.6569	0.4614

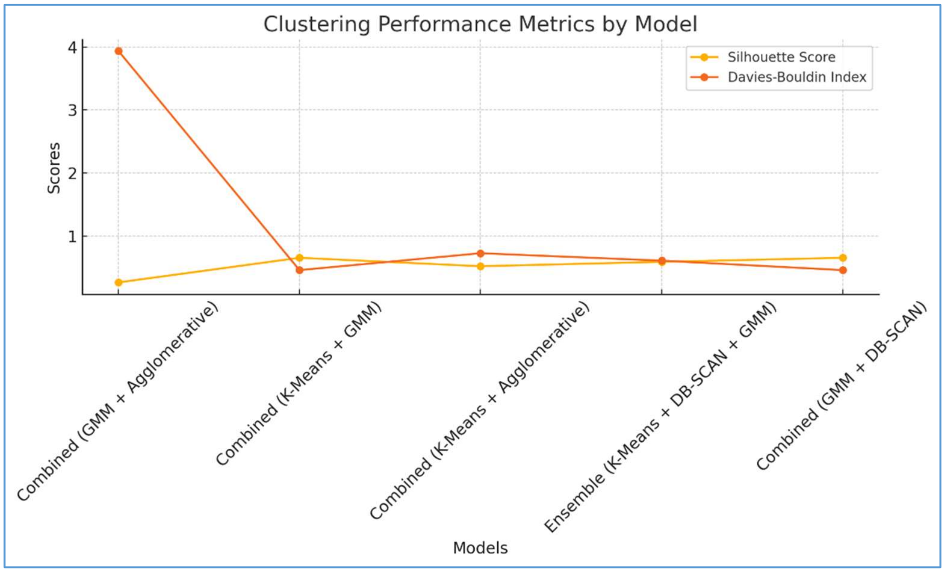


Figure 5. accuracy comparison of hybrid models

Table 3 and Figure 5 denotes that the data presented herein provides the performance of different hybrid clustering models and the comparison of the same has been done based on Silhouette Score and Davies-Bouldin Index. The best results are obtained by the Combined (K-Means + GMM) and Combined (GMM + DB-SCAN) models having both high Silhouette Score of 0.6569 and a low Davies Bouldin Index of 0.4614 suggesting well defined compact clusters. The model of Ensemble – DB-SCAN + GMM also provides fairly adequate clustering with Moderate Silhouette Score (0.5926) and Davies-Bouldin below 1 - 0.8160. Thus, the results of Combined (GMM + Agglomerative) are the lowest: Silhouette Coefficient, 0.2667, and Davies-Bouldin Index, 3.9360, prove the disturb clusters with high variation rank. Based on this we can infer that K-Means + GMM or K-Means + DB-SCAN results into the best level of clustering of the methods out of these hybrid techniques.

V. CONCLUSION

As it has been demonstrated in this paper, it is possible to incorporate the results of graph-based techniques and complex models for machine learning when designing an ITSM application in managing incident tickets. Here, we combine the monotonous multi-classification strategies including Logistic Regression, Random Forest, Gaussian Naive Bayes, XGBoost, and so on, and even introduce clustering means such as DBSCAN, and build a Gaussian Mixture Model. Experiment Result (Hierarchical Clustering), (Combined: K-Means + GMM + DB-SCAN + Agglomerative), (Combined: GMM + Agglomerative), (Combined: GMM + DB-SCAN), (Combined: K-Means + Agglomerative), (Combined: K- Means + GMM) and Ensemble (K-Means + DB-SCAN + GMM). Thus, significant increases in the standard of classifying conformance and relevancy when clustering as compared with conventional techniques. Automation and introduction of Intelligent processes–better operating efficiency and improved service quality will be developed. The flexibility and understandability of the adjunct incident ticket management systems are also enhanced. Subsequent research might be dedicated to enhancing and developing the structure of the hybrid models as well as the opportunities of the diverse graph representations for the desired system's performance.

References

[1]. M. A. Shaik, M. Parveen and I. Qureshi, “Leveraging Machine Learning and Drone Technology for Effective Insect Pest Management in Agriculture”, 2024 Asia Pacific Conference on Innovation in Technology (APCIT), MYSORE, India, 2024, pp. 1-8, doi: 10.1109/APCIT62007.2024.10673597.

[2]. M. A. Shaik, Y. Sahithi, M. Nishitha, R. Reethika, K. S. Teja and C. P. Reddy, “Realtime Emotion Recognition from Images to Understand Facial Expressions”, 2024 Asia Pacific Conference on Innovation in Technology (APCIT), MYSORE, India, 2024, pp. 1-5, doi: 10.1109/APCIT62007.2024.10673486.

[3]. Mohammed Ali Shaik, Praveen Pappula, T. Sampath Kumar, Battu Chiranjeevi, “Ensemble model based prediction of hypothyroid disease using through ML approaches”, International Conference on Research in Sciences, Engineering, and Technology, AIP Conf. Proc., 2971, 020038 (2024), https://doi.org/10.1063/5.0196055

[4]. Mohammed Ali, P. Praveen, Sampath Kumar, Sallauddin Mohmmad, M. Sruthi, “A survey report on cloud based cryptography and steganography procedures”, International Conference on Research in Sciences, Engineering, and Technology, AIP Conf. Proc. 2971, 020040-1–020040-8; https://doi.org/10.1063/5.0196050

[5]. Mohammed Ali Shaik, P. Praveen, T. Sampath Kumar, Masrath Parveen, Swetha Mucha, "Machine learning based approach for predicting house price in real estate", International Conference on Research in Sciences, Engineering, and Technology, AIP Conf. Proc. 2971, 020041-1–020041-5; https://doi.org/10.1063/5.0196051

[6]. M. A. Shaik, Y. Sahithi, M. Nishitha, R. Reethika, K. Sumanth Teja and P. Reddy, "Comparative Analysis of Emotion Classification using TF-IDF Vector," 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2023, pp. 442-447, doi: 10.1109/ICSSAS57918.2023.10331897.

[7]. M. A. Shaik, M. Azam, T. Sindhu, K. Abhilash, A. Mallala and A. Ganesh, "Hand Gesture Based Food Ordering System," 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2023, pp. 867-872, doi: 10.1109/ICSSAS57918.2023.10331637.

[8]. M. Parveen and M. A. Shaik, "Review on Penetration Testing Techniques in Cyber security," 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2023, pp. 1265-1270, doi: 10.1109/ICAISS58487.2023.10250659.

[9]. M. A. Shaik, M. Y. Sree, S. S. Vyshnavi, T. Ganesh, D. Sushmitha and N. Shreya, “Fake News Detection using NLP”, 2023 International Conference on Inovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 399-405, doi: 10.1109/ICIDCA56705.2023.10100305.

[10]. M. A. Shaik, R. Sreeja, S. Zainab, P. S. Sowmya, T. Akshay and S. Sindhu, “Improving Accuracy of Heart Disease Prediction through Machine Learning Algorithms”, 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 41-46, doi: 10.1109/ICIDCA56705.2023.10100244.

[11]. Mohammed Ali Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, “Laptop Price Prediction using Machine Learning Algorithms”, 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.

[12]. Mohammed Ali Shaik, Praveen Pappula, T Sampath Kumar, "Predicting Hypothyroid Disease using Ensemble Models through Machine Learning Approach", European Journal of Molecular & Clinical Medicine, 2022, Volume 9, Issue 7, Pages 6738-6745. https://ejmcm.com/article_21010.html

[13]. M. A. Shaik, S. k. Koppula, M. Rafiuddin and B. S. Preethi, (2022), "COVID-19

Detector Using Deep Learning", International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 443-449, doi: 10.1109/ICAAIC53929.2022.9792694.

[14]. Mohammed Ali Shaik and Dhanraj Verma, (2022), "Prediction of Heart Disease using Swarm Intelligence based Machine Learning Algorithms", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020025-1–020025-9; https://doi.org/10.1063/5.0081719, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020025-1 to 020025-9

[15]. Mohammed Ali Shaik and Dhanraj Verma, (2022), "Predicting Present Day Mobile Phone Sales using Time Series based Hybrid Prediction Model", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020073-1–020073-9; https://doi.org/10.1063/5.0081722, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020073-1 to 020073-9

[16]. Mohammed Ali Shaik, Geetha Manoharan, B Prashanth, NuneAkhil, Anumandla Akash and Thudi Raja Shekhar Reddy, (2022), "Prediction of Crop Yield using Machine Learning", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020072-1–020072-8; https://doi.org/10.1063/5.0081726, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020072-1 to 020072-8

[17]. Mohammed Ali Shaik and Dhanraj Verma, (2020), Enhanced ANN training model to smooth and time series forecast, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022038, doi.org/10.1088/1757-899X/981/2/022038

[18]. Mohammed Ali Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027

[19]. Mohammed Ali Shaik and Dhanraj Verma, (2020), Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study, 2020 IOP Conf. Ser.: Mater.Sci.Eng. 981 022041, doi.org/10.1088/1757-899X/981/2/022041

[20]. Mohammed Ali Shaik, "Time Series Forecasting using Vector quantization", International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238, Volume-29,Issue-4 (2020), Pp.169-175.

[21]. Mohammed Ali Shaik, T. Sampath Kumar, P. Praveen, R. Vijayaprakash, "Research on Multi-Agent Experiment in Clustering", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019, Pp. 1126-1129

[22]. Mohammed Ali Shaik, "A Survey on Text Classification methods through Machine Learning Methods", International Journal of Control and Automation (IJCA), ISSN:2005-4297, Volume-12,Issue-6 (2019), Pp.390-396.