MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Jamaludin *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 1068-1081

RESEARCH ARTICLE

Solving Complexity Dataset in e-Ticketing using Machine Learning to Determine Optimum Feature

Siti Zulaikha Mohd Jamaludin[a]*, Majid Khan Majahar Ali[a], Eric Vun Shiung Wong[a], Mohd Tahir Ismail[a], Noor Farizah Ibrahim[b]

[a]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Gelugor, Pulau Pinang, Malaysia; [b]School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Gelugor, Pulau Pinang, Malaysia

Abstract e-ticketing is one of the common applications used in technical support in Information Technology (IT) and has been used worldwide in any field of company. The benefits of e-ticketing can reduce the human efforts, increase the sufficiency of system and provides the benefits and efficiency to the customers. Also, e-ticketing ticketing system can enhance worker safety, improving productivity, increasing project efficiency and cause the good impact on the performance of the business in terms of profitability. The main objective of this study is defining the model performance in each important feature by analysing the complex dataset by using logistic regression as a Machine Learning (ML) algorithm. In evaluation the performance of classifier, the dataset is injected to Python programming and split into 90% as training set and 10% for the testing set. From the analysis, the study found that only 3 out of 11 independent features in dataset that are relevant chosen to proceed the ML analysis. From the result, the accuracy for sct_short_description, sct_cmdb_ci, and sct_assignment_group is 41.65%, 48.77% and 96.49%, respectively. It showed that the accuracy's result for the sct_assignment_group resulted that the model is very good accuracy and indicate that the model is well performing. Meanwhile, the value of F1-score is 96.11% in each feature. This result indicates that the model has a good balance of precision and recall in its binary classification predictions. Hence, the study considers the sct_assignment_group as a best features to proceed the analysis. The future study will consider dealing the combination of complexity features by implementing more analysis on ML such as Support Vector Machines and Naïve Bayes.

Keywords: e-Ticketing, Classification, Machine Learning, Accuracy, F1-score.

***For correspondence:**
majidkhanmajaharali@usm.my

Received: 5 March 2023
Accepted: 7 Nov. 2023

Introduction

Nowadays, ticketing system or e-ticketing has been used in the worldwide in any field of company such as transportation, entertainment, and industry. According to Gohil and Kumar (2019), e-ticketing is one of the common applications used in technical support in Information Technology (IT). Most of the support company used this system is to monitor and log all solutions and processes that come up to resolve the ticket. The IT team who has been appointed may require specific skills and expertise to solve the queries of e-ticketing accurately, as the e-Ticketing serves to connect the company and client, end to end (Yayah *et al*., 2022).

The benefit of e-ticketing can be easily access for both of clients or users at any time (Xinzhou, 2015). Abbas *et al*. (2020) and Tuveri *et al*. (2022) claimed that auto ticketing can reduce the human efforts, increase the sufficiency of system, provides the benefits and efficiency to the customers. This statement also supported by Heng and Kamsin (2021). The support team use a system to easily identify and recognize the possible issues from the client (Gohil & Kumar, 2019). With the help of the e-ticketing, it is easy for management to monitor the resolution time of the issue (Al-Hawari *et al*., 2021), keep the maintenance and appropriate scheduling for the staff (Stojanov *et al*., 2011), help the support team to point out all possible issue in the industry (Aglibar *et al*., 2022; Hojski *et al*., 2022) and it will serve as a

reference if that issue comes again on the next cycle (Robles, 2018). Also, e-ticketing ticketing system can enhance worker safety, improving productivity, increasing project efficiency and cause the good impact on the performance of the business in terms of profitability (Utama *et al*., 2021; Robertson *et al*., 2022; Tripathi *et al*., 2022).

In purpose to discover the significant motivations and factors that influence the quality of e-ticketing towards customers' perception, preferences and intentions in e-commerce business, Smith *et al*. (2014) has conducted a survey with 39 item questions in nominal scale. The study applied statistical analysis of multiple linear regression and found that the customers' perception, preferences and intentions gives significant positive relationship to promoting the perceived value of e-ticketing technologies. Sun (2016) has conducted an online survey to investigate the needed of customer towards the implementation of e-ticketing in subway transportation and found that the e-ticketing can save of time passenger from queue up for tickets. Aglibar *et al*. (2022) used data collected from the service now open-source (SNOW) ticketing system and performed a survey from management and support team to analyze the productivity of the staff in industry's sector by using a e-ticketing. They found that critical and auto ticket's affected daily productivity of the staff.

Agarwal *et al*. (2012) has demonstrated the efficiency of dispatching the ticket from the system by gathered data from SmartDispatch. They used supervised learning technique; Support Vector Machine (SVM) and Discriminative Term Approach (DTA), to perform the classification model that uses ticket text description to predict resolution group. They implemented techniques with SPSS software and found that the DTA is more perform in dispacthing the ticket to the correct resolution group with low error rates. Yayah *et al*. (2022) found that most of study that gathered the data from system usually has implemented data preprocessing method since the study dealing with big data. The study used data from Apache Sqoop system to track the detection, reporting and resolution of tickets submitted by telco customers. They approached single machine (traditional classifier) and and Hadoop (advanced classifier) to solve the of record limits and improved classification accuracy. They found that the accuracy classification of Hadoop improved approximately 8% compared to single machine.

In 2018, Qamili *et al*. used primary data and proposed machine learning to threat the spam detection, ticket assignment and sentiment analysis in e-ticketing. They found Random Forest is the best model perform in threating spam detection (ACC = 99.85%) and ticket assignment (ACC = 86.1%). Al-Hawari *et al*. (2021) has applied supervised machine learning to perform the classification in help desk system. They mentioned the steps to build an accurate and fast ticket classification model and found the accuracy of model prediction increased from 53.8% to 81.4%.

The Figure 1 illustrate the use case of problem that address in XYZ Company IT Services, where the invalid assignment of groups done by both IT expertise and non-expertise, where whole IT services management still looking for solution to prevent this problem. Noted that this study is focused mainly only WEB-Based tickets which logged by users or non-expertise, where the assignment groups manually filled up by users themselves due to the readily available of dataset on this medium compared to other mediums and to have textual data that explain the IT issues in different ways to have deep data analysis using ML techniques and tools.

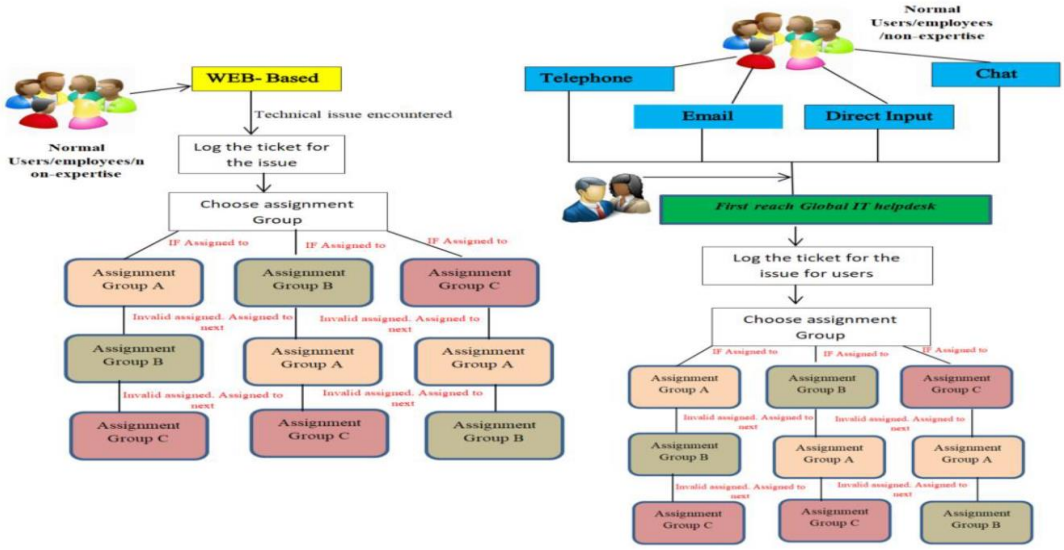

Figure 1. Illustration of use case problem

Numerous studies have shown that there are some drawbacks in system such as a ticket features consists of a short of description and fields capturing category-specific information about ticket from the client (Gupta *et al*., 2018). Due to this, it causes difficulties to the IT team to analyze and recognize the problem in traditional methods. With this limitation, it could lead the IT services wrong assigning the ticket to the right team, increasing the cost for support providing company and stretching the resolution time. According to Paramesh *et al*. (2019) the crucial step in the problem resolution process is speedy dispatching the ticket to the correct resolution group while maintaining the level of high accuracy, especially in traditional way of ticketing.

Table 1 shows the summary of previous studies on e-ticketing system. From the table, the first gap that we found from the previous studies is not a lot of studies do not apply preprocessing data, especially in categorical attibutes. Second gap is most of previous studies only focusing on simulation study on other conventional methods compared to the machine learning. Third gap is most of studies which it applied data cleaning did not mentioned clearly on how they handle missing value and null string value in dataset. This problem will caused the complication in do analysis and to proceed the further analysis. Also, it is impossible to analyse complex dataset using traditional statistical method.

Table 1. Previous studies on ticketing system

Author(s) (Year)	Type of Paper		Application of Data		Data Preprocessing		Type of Analysis			Tools of Analysis		Remark
	Review	Research	Industry	Others	Yes	No	Statistical Model	Machine Learning	Hybrid Model	Software	Programming	
Qamili *et al*. (2018)		√	√		√		√			√	√	Software company.
Smith *et al*. (2014)		√		√		√	√	√		√		Entertainment and sport business.
Al-Hawari *et al*. (2021)		√		√	√		√	√			√	Help desk system – MyGJU portal.
Abbas *et al*. (2020)		√		√								Bus & train.
Sun (2016)		√		√								Subway ticket.
Tuveri *et al*. (2022)		√		√								Public transport.
Yayah *et al*. (2022)		√		√	√		√					Telco company.
Agarwal *et al*. (2012)		√		√	√		√				√	IBM Company.

Thus, the first objective of this study will be focus on selecting the important feature in e-ticketing rergarding CTask for XYZ Company IT services department in solving the limitation. Second objective will discuss on defining the model performance in each important variable by analyzing the complex dataset by using ML.

The remaining part of this paper is as follows: Section 1 explores the uses of machine learning in e-ticketing system, especially in c-tasking dataset; Section 2 explains the material and method in detail; Section 3 discusses the result and discussion thoroughly and Section 4 gives the concluding remarks.

Methodology

With the aim to achieve the objective in selecting important features and evaluate model performance, the work needs to consider all the features that containing the heterogeneous textual or categorical. It is necessary to generate a numeric form of description that can be consumed by classification methods (Agarwal *et al*., 2012). The process of the data analysis setup is shown in Figure 2.

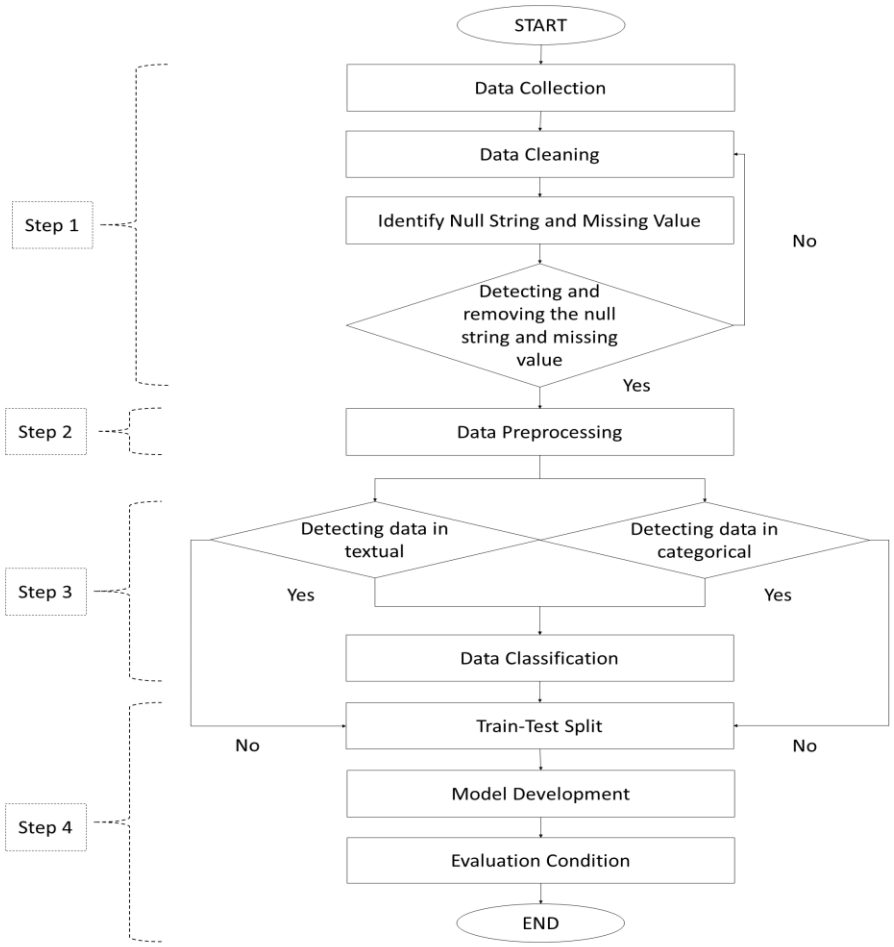

Figure 2. Flowchart of the data analysis setup

During the process of the data analysis setup, the data had injected into Python programming to determine the details of process as discussed as follows:

Data Collection

The data used is Ctask dataset which has been extracted from the e-ticketing of a XYZ company. Initially, the dataset consists of 37532 records with 12 features in total are generates from customer report from which the tickets are automatically generated. 11 indepedent features that used in this work are sct_number, sct_short_description, sct_cmdb_ci, sct_state, sct_assignment_group, sct_assignment_to, sct_escalation, mi_definition, mi_value, mi_sys_created_on and sct_closed_at. Meanwhile the dependent features is sct_priority. The summary of features with their respective description as shown in the following Table 2.

Table 2. The features and description of Ctask dataset

Features	Description	Type of Features
sct_number	The ticket number (unique), which will be automatically generated by the system.	String
sct_short_description	Description of the technical issue mentioned by the customer.	String
sct_cmdb_ci	Known as Configuration Item, that specify particular issue (example: Computer issue).	Categorical

Features	Description	Type of Features
sct_priority	Urgency level based on user preference (low, medium, high or critical).	Categorical
sct_state	Current state of ticket (incomplete, skipped or complete).	Categorical
sct_assignment_group	The IT team who resolved this issue/ticket.	Categorical
sct_assignment_to	The staff who resolved this issue/ticket.	Categorical
sct_escalation	The type of current escalation of issue.	Categorical
mi_definition	The definition of current issue.	Categorical
mi_value	The staff who closed the issue/ticket.	Categorical
mi_sys_created_on	Time when the ticket of issue created (timestamp).	Time series
sct_closed_at	Time when the ticket of issue done (timestamp).	Time series

Data Cleaning

Note that, the unimportant features may cause detrimental effect on the model's performance. A crucial idea that affects machine learning is the selecting of features during data cleaning (Yayah *et al*., 2022). At early this stage, the dataset was injected into Microsoft Office Excel to view the dataset accordingly in the form of rows and columns to perform some data cleaning. Removal of irrelevant column or feature from the dataset is the initial step of data cleaning process, where there are 8 columns including sct_number, sct_state, sct_assignment_to, sct_escalation, mi_definition, mi_value, mi_sys_created_on and sct_closed_at have been removed completely. The reasons of removing those columns are presented in the below Table 3.

Table 3. Reason for removal of columns

Features	Reason to Removed
sct_number	Removed. No patterns found. The features for ticket unique reference, could not be able to analyse and visualize the instances within this feature.
sct_state	Removed. The overall chosen state was “Closed Complete”, where only resolved and closed tickets are obtained from the repository in order to trained the model, due to the completeness of issue by the respective team.
sct_assignment_to	Removed. No irrelevant found for data analysis. No patterns found out and not visually provide any information of work.
sct_escalation	Removed. The overall chosen state was “Normal”. No patterns found out and not visually provide any information of work.
mi_definition	Removed. The overall chosen state was “Catalog Task Assigned to Duration”. No patterns found out and not visually provide any information of work.
mi_value	Removed. No irrelevant found for data analysis. No patterns found out and not visually provide any information of work.
mi_sys_created_on	Removed. Time is not relevant to the scope of work.
sct_closed_at	Removed. Time is not relevant to the scope of work.

After finalizing the important features, it is important to handle missing values properly before conducting any analysis. The benefit this is for preventing invalid models coming from the values that are missing or null (Yayah *et al*, 2022). Missing values in a dataset can have several effects such as minimize the quality and quantity of classification (Al Shalabi, 2016), lead to difficulties extracting the important information from the dataset (Li & Sharma, 2022) and bias in the results of the analysis (Penone *et al*., 2014).

With this, the `pandas` module is imported in python programming. With the aims of detect the total missing value and null string in each feature, isna function will be applied. Afterwards, the function of dropna will be employed used for removing all that missing value and null string. The pseudocode of this process is as follows:

```
import pandas as pd
data.isna().sum()
data["variable1"].iloc[r1:r2]
data.dropna(inplace=True)
```

Data Preprocessing

The process of preparing the raw dataset for analysis need considerable work. After done with data cleaning, `sklearn.preprocessing` and `nltk.tokenize` is applied to proceed data preprocessing for categorical and textual feature, respectively. The following list contains detailed explanation on how to perform the data preprocessing in each type of features:

1. Categorical Feature

As the dataset is consisted categorical feature, the `LabelEncoder` is imported into python. Emphasize that, the categorical feature must be encoded to numbers before fit to the model and perform the evaluation in ML. This is because some datasets need to change to the correct format in purpose to meet the need of the model of ML. The pseudocode of label encoder is as follows:

```
from sklearn.preprocessing import LabelEncoder
data=LabelEncoder()
```

2. Textual Feature

Initially, text cleaning methods is applied in data preprocessing for textual feature. The NLTK package is applied in purpose to remove of any stop words or punctuation in data that have very little meaning or are irrelevant in the work. During this stage, the larger body of text (paragraph) is split into words, keywords and symbols called “tokens,” which are delimited by commas or white spaces. Thus, `stopwords` and `string` imported along with a custom python script and used for removal of those words in natural language.

Also, this work employed a bag-of-words approach meaning that individual words correspond to features for further processing (Yayah *et al*, 2022)., which are obtained using a `TweetTokenizer` and `word_tokenize`. The pseudocode of data tokenization is as follows:

```
import nltk
from nltk.tokenize import TweetTokenizer
from nltk.tokenize import word_tokenize
tokenizer = TweetTokenizer(preserve_case=True,
    strip_handles=False,
    reduce_len=False, match_phone_numbers=True)
data['tokenize'] = data["variable1"].apply(tokenizer.tokenize)
```

To covert the inflected word to its root form, the lemmatization technique is applied. With that, WordNetLemmatizer and wordnet is imported along with a custom python script. It is necessary in order to group similar words together after the removal of stop words from the data. For example, big deksttop - > desktop. The pseudocode of lemmatization is as follows:

```
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet
def get_wordnet_pos(treebank_tag):
    if treebank_tag.startswith('J'):
       return wordnet.ADJ
    elif treebank_tag.startswith('V'):
        return wordnet.VERB
    elif treebank_tag.startswith('N'):
        return wordnet.NOUN
    elif treebank_tag.startswith('R'):
        return wordnet.ADV
    else:
        return None
lemmatizer = WordNetLemmatizer()
```

Data Classification

Next, the exploratory process further preceded by convert the processed textual feature into numerical form in purpose to train the models using ML algorithms. The TF-IDF Vectorizer will be apply in this study

to transform text to feature vectors that are used as input for further machine learning approaches (Qorib *et al*. 2023). According to Malviya, & Dwivedi (2022), TF-IDF is preferable used to distinguish very common or rare words and gives a measure that takes the importance of word into consideration depending on how frequently it occurs in a document and a corpus. This statement is supported by Jayasurya *et al*. (2021).

"TF" is the "count of the words presents in the document from its own vocabulary". Also, it defined as the percentage of the number of times a word (x) occurs in a particular document (y) divided by the total number of words in that document:

$$\text{TF(Term)}=\frac{\text{Number of times term appears in a document}}{\text{Total number of items in the document}} \quad (1)$$

While "idf" is the "importance of the word to each document" or the logarithmic ratio of no. of total documents to no. of a document with a particular word:

$$\text{IDF(Term)}=\log\left(\frac{\text{Number of times term appears in a document}}{\text{Total number of items in the document}}\right) \quad (2)$$

To convert the text data into numerical form via the TF-IDF, `sklearn` library has been injected into python and the pseudocode is as follows:

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_model = TfidfVectorizer(binary=True)
X_train_tfidf = tfidf_model.fit_transform(X_train).astype('float16')
X_test_tfidf = tfidf_model.transform(X_test).astype('float16')
```

Model Development

This study focusses on the he ML algorithm of logistic regression (LR) to performed to select the feature. LR is applied due to systematically ranked among the best models in ML as it is widely used ML algorithm that is well-suited for binary classification problems (Nusinovici *et al*., 2020). Additionally, LR also has relatively low computational complexity, which makes it a good choice for problems with large datasets (Hassan & Ali, 2018). Also, Dedeturk and Akay (2020) has mentioned that LR also can handle high-dimensional feature spaces, where the number of features is much larger than the number of instances. In this study, LR is used to model the single relationship on between single each selected independent features and dependent feature. The formula of LR is in the form as follows:

$$l\left(\frac{P}{1-P}\right)=\alpha_0+\alpha_1X_1+\ldots+\alpha_nX_n$$

$$\frac{P}{1-P}=e^{\alpha_0+\alpha_1X_1+\ldots+\alpha_nX_n} \quad (3)$$

$$P=\frac{e^{\alpha_0+\alpha_1X_1+\ldots+\alpha_nX_n}}{1+e^{\alpha_0+\alpha_1X_1+\ldots+\alpha_nX_n}}$$

In python programming, `LogisticRegression` has injected to perform the LR analysis. The pseudocode is as follows:

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(random_state=0).fit(X_train_tfidf, y_train)
y_pred = model.predict(X_test_tfidf)
```

Evaluation Condition

The study follows the state-of-the-art of Vabalas *et al*. (2019) and Kalaycioglu *et al*. (2022) to evaluates the performance of the model on dataset by splitting the training dataset into 90% and testing dataset into 10%. The work used the `model_selection` module from scikit-learn library, in which we have the splitter function `train_test_split()`. This study used accuracy (ACC) and F1-score to evaluate the model performance. (Chola *et al*., 2022). The use of ACC is to define the percentage of model performance in correct prediction (Sigh & Kumar, 2020). Also, ACC is used to evaluate how well the model has been trained on a given dataset (Poczeta *et al*., 2023). As state by Gumilar *et al*. (2021), F1-score is used to measure how much the model able to predict the class correctly. Additionally, F1-score is the harmonic mean of precision and recall, where a higher F1-score indicates a better balance between precision and recall (Zangari *et al.,* 2023; Kasihmuddin *et al.,* 2023). In this study, the ACC will portray on how well estimation the performance of model on a given dataset and F1-score will portray on evaluation the ability of the model to accurately classify instances in a binary classification problem. The

formula of ACC dan F1-Score as follows in Equation (4) and (5). Note that, True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are derived via a confusion matrix.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$F1 = 2 \times \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}} \quad (5)$$

The phyton's pseudocode of performance model is as follows:

```
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix
accuracy_score(y_test, y_pred)
f1_score(y_test, y_pred, average='weighted')
confusion_matrix(y_test, y_pred)
```

Results and Discussion

The outcome of the columns removal actions in data cleaning process resulted in only 4 columns of features that are relevant chosen includes sct_assignment_group, sct_cmdb_ci, sct_short_description and sct_priority, as shown in Table 4. All those features contain the instances that are heterogeneous textual and categorical in form was further injected into Python to determine the missing values using Python's codes (Singh & Kumar, 2020).

Table 4. The features and description of Ctask dataset

Features	Description	Type of Features
sct_assignment_group	The IT team who resolved this issue/ticket.	Categorical
sct_cmdb_ci	Known as Configuration Item, that specify particular issue (example: Computer issue).	Categorical
sct_short_description	Description of the technical issue mentioned by the customer.	Textual
sct_priority	Urgency level based on user preference (low, medium, high or critical).	Categorical

From the Python's result, the study found that there is missing value and null string in sct_short_description and sct_cmdb_ci. While there is no missing values and null strings in sct_priority and sct_assignment_group. The count of missing value and null strings in features is shown in Table 5. Due to number of missing value and null string is very low count, the study decided to eliminate them from dataset to perform data preprocessing and this is supported by study of Chang *et al*. (2021).

Table 5. Counts of missing value and null string

Features	Count
sct_assignment_group	0
sct_cmdb_ci	11
sct_short_description	7
sct_priority	0

During the exploratory process, further extended by visualizing the initial counts on sct_assignment_group and sct_cmdb_ci using Python's code. The early analysis has resulted that the sct_assignment_group has 302 initial counts and the sct_cmdb_ci has 527 initial counts. The visual of initial count is representation using bar chart, and the output is shown in the below Figure 3 and Figure 4.

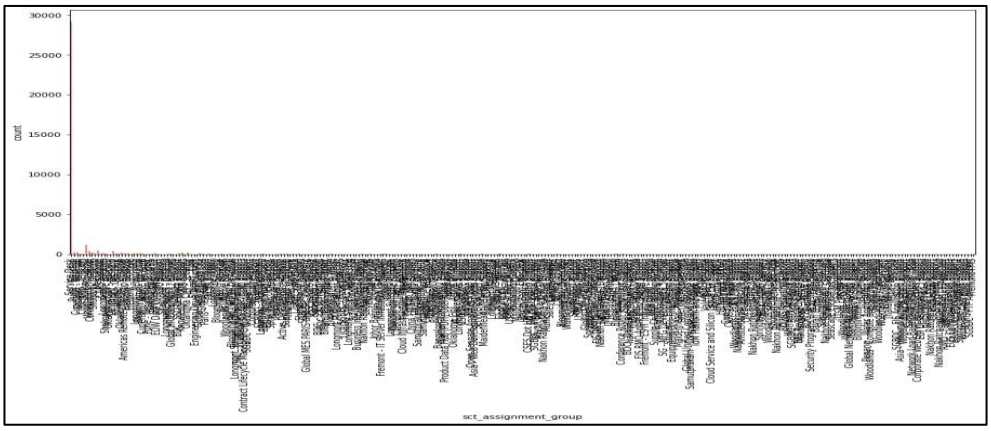

Figure 3. Initial visualization of sct_assignment_group

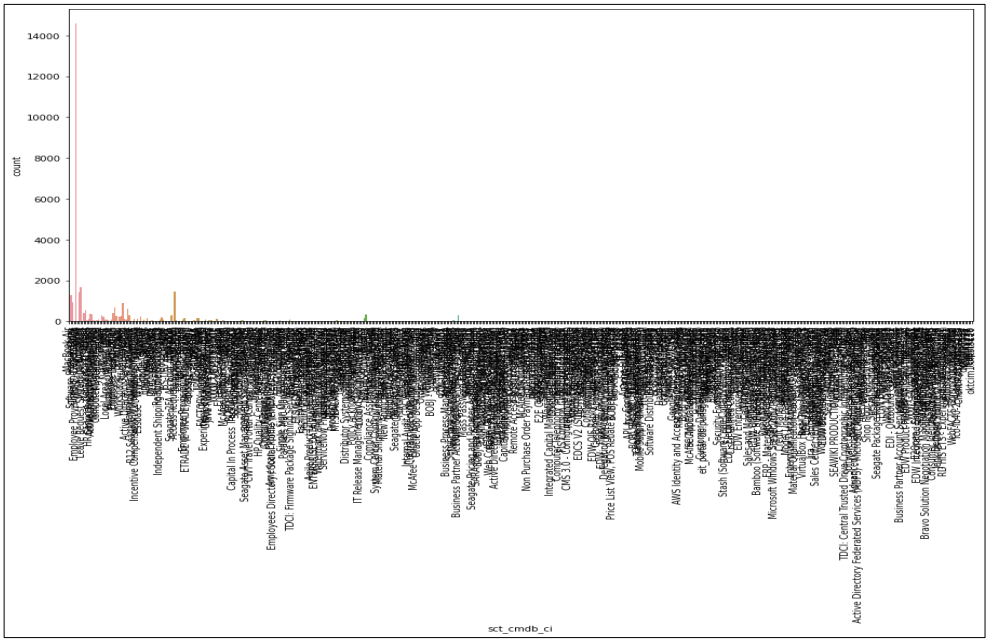

Figure 4. Initial visualization of sct_cmdb_ci

The output shows, there huge difference found between the sct_assignment_group and sct_cmdb_ci counts that shows imbalance of classes, therefore this may cause the data analysis process to be complex by encountering biased values. Therefore, the dataset been undergoing cleaning process by removing least number of sct_assignment_group for instance technical issue and sct_cmdb_ci for instance particular issue to balance the classes. The cleaning process further proceeded by merging the multiple assignment groups into one big group. The removing and merging the features resulted in only 5 group in each sct_assignment_group and sct_cmdb_ci to be finalized for the data analysis process, as shown in below Figure 5 and Figure 6.

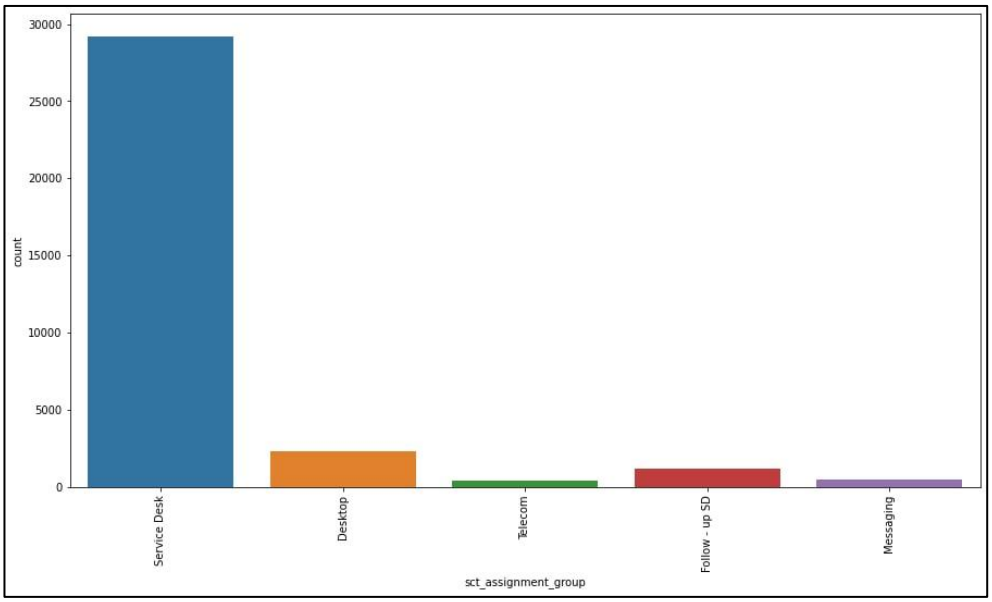

Figure 5. Final visualization of sct_assingment_group

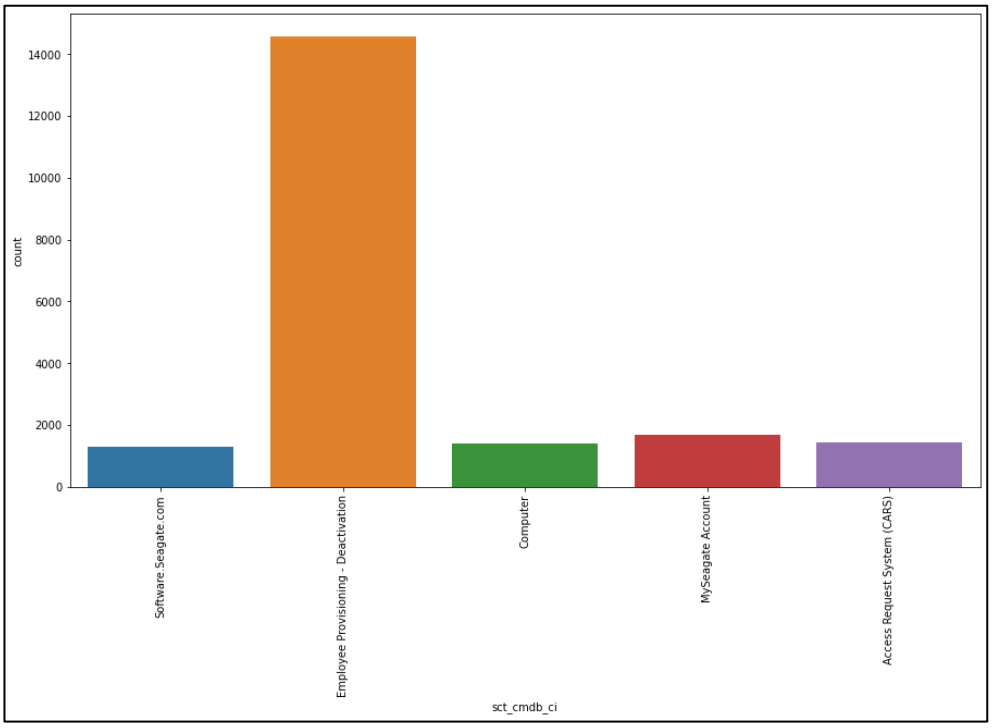

Figure 6. Final visualization of sct_cmdb_ci

The finalized dataset contains 33655 rows of sct_assingment_group with 5 labelled of classes of Service Desk, Desktop, Telecom, Follow-up SD and Messaging as shown in Table 6. While the finalized dataset contains 21279 rows of sct_cmdb_ci with 5 labelled of classes of Employee Provisioning - Deactivation, MySeagate Account, Access Request System (CARS), Computer and Software.Seagate.com as shown in Table 7.

Since the finalized dataset showed in categorical form, the study then proceeds encoded dataset to the numbers. The result of label encoded for class label in each feature is shown as in Table 6 and 7.

Table 6. Label Encoder for sct_assingment_group

Class Label	Final Count	Numerical Label
Service Desk	29195	3
Desktop	2353	0
Follow-up SD	1221	1
Messaging	478	2
Telecom	408	4

Table 7. Label Encoder for sct_cmdb_ci

Class Label	Final Count	Numerical Label
Employee Provisioning - Deactivation	14588	2
MySeagate Account	1682	3
Access Request System	1452	0
Computer	1408	1
Software.Seagate.com	1284	4

Initially, the results of visualizing the initial counts on sct_priority shows the balance clasess as in Figure 7. Thus, does no need to undergoes cleaning process by removing least number. However, this feature had performed the numerical transformation to proceed the further statistical analysis and the result is shown in Table 8.

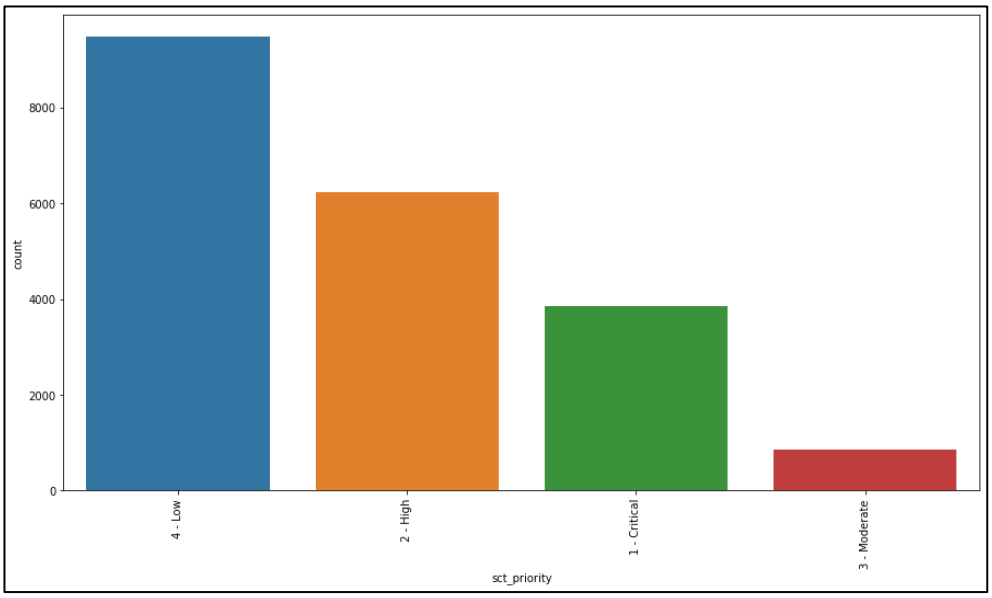

Figure 6. Visualization of sct_priority

Table 8. Label Encoder for sct_priority

Class Label	Final Count	Numerical Label
4 – Low	9475	3
2 – High	6226	1
1 – Critical	3857	0
3 – Modetrate	856	2
4 – Low	9475	3

Then, the dataset preprocess for further extended by removing stop words, transforming words to lower letters, removing punctuation, and lemmatization using NLTK tokenization in Python for textual feature. For this study, a word tokenize module used for NLTK method to process the feature of sct_short_description.

In evaluation the performance of classifier, the 30304 of dataset is split into 90% as training set and 10% for the testing set. This study chose logistic regression as an algorithm provided by scikit-learn that is recommended when dealing with supervised learning and as well in document classification problems (Schade & Schuhmacher, 2023; Zheng *et al*., 2021). The result of the performance of model algorithm for each feature is shown in Table 9.

Table 8. Model performance of ACC and F1-score for each independent feature

Feature	Model Performance	
	ACC (%)	F1-Score
sct_assignment_group	41.65	96.11
sct_cmdb_ci	48.77	96.11
sct_short_description	96.49	96.11

Based on the performance measures output, the accuracy of logistic regression algorithm for sct_assignment_group, sct_cmdb_ci and sct_short_description is 41.65%, 48.77% and 96.49, respectively. The accuracy's result for the sct_short_description resulted that the model is very good accuracy and indicate that the model is well performing (Chowdhury *et al.,* 2023). However, unlikely the accuracy's result for the sct_assignment_group and sct_cmdb_ci. Both accuracy's value showed low accuracy and indicate that those both models not well performing. This is might due to variety of factors such as overfitting, underfitting, or poor feature selection.

The value of F1-score of each features showed same with value of 96.11%. This value indicates that the model has a good balance of precision and recall in its classification predictions (DeVries *et al*., 2021). Thus, it can be concluded that the best feature in this study is sct_short_description, followed by sct_cmdb_ci and sct_assignment_group.

Conclusions

In outline, only 3 out of 11 independent features in Ctask dataset that are relevant chosen to proceed the ML analysis, which are includes sct_short_description, sct_cmdb_ci, and sct_assignment_group. While the dependent feature is sct_priority. The study implements logistic regression analysis as ML algorithm to analyse the complex dataset and define the model performance in each important feature. From the result and analysis, the feature of sct_assignment_group showed the higher accuracy (96.49%) and F1-score (96.11%). Thus, it considers the feature as a best feature to proceed the analysis. Since the data e-ticketing is in binary the classification prediction, the value of F1-score is more applicable to indicates that the model has a good balance of precision and recall. Indirectly, it would help the company to automate the right group to the level of priority to ensure timely responses, efficient manner and resolutions when they in resolve the e-ticketing problem.

This study only addressed the one ML algorithm which is LR by analyse in each single complex feature. This is due to the limitation of LR does not perform well when the relationship between the input features and the target feature is highly non-linear, or when there are complex interactions between the features (Mordensky *et al*., 2023). As a part of future work, the study will consider dealing the combination of complexity features by implementing more analysis on ML such as Support Vector Machines (Mayes *et al*., 2023) and Naïve Bayes (Sawhney *et al*., 2023), k-Nearest Neighbor (Singh *et al*., 2022) and Random Forest (Zafeiropoulos et al., 2023) in solving the problem of e-ticketing system.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

This work is supported by the School of Mathematical Sciences, Universiti Sains Malaysia.

References

[1] Aglibar, K. D., & Rodelas, N. (2022). Impact of Critical and Auto Ticket: Analysis for Management and Workers Productivity in using a Ticketing System. *International Journal of Computing Sciences Research*, *6*, 988-1004.
[2] Al Shalabi, L., Najjar, M., & Al Kayed, A. (2006). A framework to deal with missing data in data sets. *Journal of Computer Science*, *2(9)*, 740-745.
[3] Al-Hawari, F., & Barham, H. (2021). A machine learning based help desk system for IT service management. *Journal of King Saud University-Computer and Information Sciences*, *33(6)*, 702-718.
[4] Bhanot, N., Ahuja, J., Kidwai, H. I., Nayan, A., & Bhatti, R. S. (2022). A sustainable economic revival plan for post-COVID-19 using machine learning approach–a case study in developing economy context. *Benchmarking: An International Journal*, (ahead-of-print).
[5] Chang, W., Ji, X., Wang, L., Liu, H., Zhang, Y., Chen, B., & Zhou, S. (2021, October). A Machine-Learning method of predicting vital capacity plateau value for ventilatory pump failure based on data mining. In Healthcare (Vol. 9, No. 10, p. 1306). Multidisciplinary Digital Publishing Institute.
[6] Chola, C., Muaad, A. Y., Bin Heyat, M. B., Benifa, J. V., Naji, W. R., Hemachandran, K., ... & Kim, T. S. (2022). BCNet: A Deep Learning Computer-Aided Diagnosis Framework for Human Peripheral Blood Cell Identification. *Diagnostics*, *12(11)*, 2815.
[7] Chowdhury, M. Z. I., Leung, A. A., Walker, R. L., Sikdar, K. C., O'Beirne, M., Quan, H., & Turin, T. C. (2023). A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a Canadian population. *Scientific Reports*, *13(1)*, 13.
[8] Dedeturk, B. K., & Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing*, *91*, 106229.
[9] DeVries, Z., Locke, E., Hoda, M., Moravek, D., Phan, K., Stratton, A., ... & Phan, P. (2021). Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability. *The Spine Journal*, *21(7)*, 1135-1142.
[10] Gohil, F., & Kumar, M. V. (2019). Ticketing system. International Journal of Trend in Scientific Research and *Development*, *3(4)*, 155-156.
[11] Gupta, M., Asadullah, A., Padmanabhuni, S., & Serebrenik, A. (2018). Reducing user input requests to improve IT support ticket resolution process. *Empirical Software Engineering*, *23(3)*, 1664-1703.
[12] Hassan, M. A., & Ali, A. (2018). Logistic regression and gradient descent algorithms for big data classification. *Journal of Big Data*, *5(1)*, 1-19.
[13] Hojski, D., Hazemali, D., & Lep, M. (2022). The Analysis of the Effects of a Fare Free Public Transport Travel Demand Based on E-Ticketing. *Sustainability*, *14(10)*, 5878.
[14] Kasihmuddin, M. S. M., Jamaludin, S. Z. M., Mansor, M. A., Wahab, H. A., & Ghadzi, S. M. S. (2022). Supervised learning perspective in logic mining. Mathematics, *10*(6), 915.
[15] Li, F., & Sharma, A. (2022). Missing Data Filling Algorithm for Big Data-Based Map-Reduce Technology. *International Journal of e-Collaboration (IJeC)*, *18(2)*, 1-11.
[16] Malviya, A., & Dwivedi, R. K. (2022). Detecting Deceptive News in Social Media Using Supervised Machine Learning Techniques. In IOT with Smart Systems: Proceedings of ICTIS 2022, Volume 2 (pp. 239-250). Singapore: Springer Nature Singapore.
[17] Mayes, E., Gehlbach, J. A., Jeziorczak, P. M., & Wooldridge, A. R. (2023). Machine Learning to Operationalize Team Cognition: A Case Study of Patient Handoffs. *Human Factors in Healthcare*, 100036.
[18] Mordensky, S. P., Lipor, J. J., DeAngelo, J., Burns, E. R., & Lindsey, C. R. (2023). When less is more: How increasing the complexity of machine learning strategies for geothermal energy assessments may not lead toward better estimates. *Geothermics*, *110*, 102662.
[19] Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56-69.
[20] Paramesh, S. P., & Shreedhara, K. S. (2019). Automated IT service desk systems using machine learning techniques. In Data Analytics and Learning (pp. 331-346). Springer, Singapore.
[21] Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., ... & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best?. *Methods in Ecology and Evolution*, *5(9)*, 961-970.
[22] Poczeta, K., Płaza, M., Michno, T., Krechowicz, M., & Zawadzki, M. (2023). A multi-label text message classification method designed for applications in call/contact centre systems. Applied Soft Computing, 110562.
[23] Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, *212*, 118715.
[24] Robertson, G., Zhang, S., & Bogus, S. M. (2022). Challenges of Implementing E-Ticketing for Rural Transportation Construction Projects. In Construction Research Congress 2022 (pp. 453-462).
[25] Rusli, A., Suryadibrata, A., Nusantara, S. B., & Young, J. C. (2020). A Comparison of Traditional Machine Learning Approaches for Supervised Feedback Classification in Bahasa Indonesia. *International Journal of New Media Technology*, *7(1)*, 28-32.
[26] Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment of artificial intelligence

models used for early prediction and evaluation of chronic kidney disease. *Decision Analytics Journal*, 100169.
[27] Schade, P., & Schuhmacher, M. C. (2023). Predicting entrepreneurial activity using machine learning. *Journal of Business Venturing Insights*, *19*, e00357.
[28] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.
[29] Singh, H., Sharma, V., & Singh, D. (2022). Comparative analysis of proficiencies of various textures and geometric features in breast mass classification using k-nearest neighbor. *Visual Computing for Industry, Biomedicine, and Art*, *5*, 1-19.
[30] Stojanov, Z., Dobrilovic, D., & Jevtic, V. (2011, September). Identifying properties of software change request process: Qualitative investigation in very small software companies. In 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics (pp. 47-52). IEEE.
[31] Tripathi, A., Patel, D., Sturgill, R., & Dadi, G. B. (2022). Analysis of E-Ticketing Technology for Inspection Performance and Practicality on Asphalt Paving Operations. Transportation Research Record: 03611981221083308.
[32] Utama, A. A. G. S., Astuti, P. P. D., Hikmawati, E. E., & Setyowati, Y. (2021, November). Design E-Ticketing System to Increase Ticket Sales in Banyuwangi Branch New Star Cineplex. In International Conference on Management, Business, and Technology (ICOMBEST 2021) (pp. 120-126). Atlantis Press.
[33] Xinzhou, H. (2015). The development trend of mobile Internet and its impact on the global economy. In 2015 International Conference on Social Science and Technology Education (pp. 598-600). Atlantis Press.
[34] Zafeiropoulos, N., Mavrogiorgou, A., Kleftakis, S., Mavrogiorgos, K., Kiourtis, A., & Kyriazis, D. (2023). Interpretable Stroke Risk Prediction Using Machine Learning Algorithms. In Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2 (pp. 647-656). Singapore: Springer Nature Singapore.
[35] Zangari, A., Marcuzzo, M., Schiavinato, M., Gasparetto, A., & Albarelli, A. (2023). Ticket automation: An insight into current research with applications to multi-level classification scenarios. *Expert Systems with Applications*, 119984.
[36] Zheng, L., Wen, L., Lei, W., & Ning, Z. (2021). Added value of systemic inflammation markers in predicting pulmonary infection in stroke patients: A retrospective study by machine learning analysis. *Medicine*, *100(52)*.