

---

# Journal of Informatics and Web Engineering

Vol. 5 No. 3 (October 2026)

eISSN: 2821-370X

---

## Structured Data Preprocessing Techniques for a Multilingual Healthcare Dataset in Machine Learning Applications

**Md. Abu Nishan<sup>1</sup>, Choo-Yee Ting<sup>2\*</sup>, Chih-Yang Pee<sup>3</sup>, Helmi Zakariah<sup>4</sup>, Thanh-Trung Nguyen<sup>5\*\*</sup>**

<sup>1,2,3</sup>Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia

<sup>4</sup>Hayat Technologies Sdn Bhd, Level 15, Centerpoint North, Lingkaran Syed Putra, Mid Valley City, 59200 Kuala Lumpur, Malaysia

<sup>5</sup>Faculty of Information Technology - Communications, Dong Nai Technology University, Dong Nai, Vietnam

\*corresponding author: (cyting@mmu.edu.my; ORCID: 0000-0001-5667-2816)

\*\*corresponding author: (nguyenthanhtrung@dmu.edu.vn; ORCID: 0009-0000-9948-1753)

**Abstract** – Data preprocessing is a foundational stage in any machine learning (ML) workflow, especially when dealing with complex, multilingual, and real-world healthcare datasets. The quality, consistency, and structure of input data strongly influence the accuracy, fairness, and interpretability of predictive models. This paper presents a comprehensive and structured preprocessing framework designed for a large multilingual healthcare dataset containing 19,536 patient records with 67 features encompassing demographic, clinical, and socioeconomic information. The framework systematically addresses data heterogeneity, linguistic diversity, and structural inconsistencies through multiple stages of cleaning and transformation. The preprocessing process includes translation of Malay text into English to ensure linguistic uniformity, standardized handling of missing values using transparent placeholders, and column splitting of composite attributes to improve granularity and feature representation. Outlier detection was performed using statistical approaches such as Interquartile Range (IQR) and Z-Score methods to identify and remove extreme or erroneous values, thereby improving data integrity and model stability. Furthermore, categorical variables were converted into numerical form using one-hot encoding, ensuring compatibility with diverse ML algorithms while preventing artificial ordering and bias. An experimental evaluation using a Random Forest classifier demonstrated that these preprocessing techniques substantially enhanced predictive performance. Model accuracy improved from 72.4% to 86.7%, while the area under the curve (AUC) increased from 0.78 to 0.92, indicating more robust classification and generalization. The resulting dataset is cleaner, more consistent, and better suited for downstream ML applications. Overall, this study provides a reproducible and practical framework for managing multilingual healthcare data, underscoring that systematic preprocessing is not merely a preliminary task but a critical foundation for building reliable, transparent, and interpretable machine learning solutions in healthcare and other data-intensive domains.

**Keywords**— Data Preprocessing, Multilingual Healthcare Data, Data Cleaning, Outlier Detection, Translation, One-Hot Encoding, Machine Learning.

*Received: 01 July 2025; Accepted: 01 January 2026; Published: 16 October 2026*

*This is an open access article under the [CC BY-NC-ND 4.0](#) license.*



## 1. INTRODUCTION

In the world of machine learning, data is everything, but raw data is rarely ready for analysis. It is often messy, inconsistent, incomplete, and full of errors or unclear parts. This is where data preprocessing steps come into play. It is a key but often less discussed stage that changes chaotic real-world data into a structured, useful, and machine-readable format.

Preprocessing is not just about cleaning up data. It is a planned process that involves cleaning, changing, and organizing data to help it reach its full value [1]. When done well, it improves data quality and makes sure learning algorithms focus on real patterns instead of noise or random mistakes [2]. Even though it is important, preprocessing often gets less attention than model choice or parameter tuning. However, poor preprocessing can ruin even the best models.

The value of preprocessing is in its power to build a strong base for model success. Without it, algorithms may learn wrong trends, strengthen unfair biases, or give random results. Think of a healthcare dataset with missing values, different formats, or wrong labels. Using such raw data in a model can give bad generalization, wrong predictions, or even harmful real-world decisions. Preprocessing helps fix these problems by correcting mistakes, handling missing values, finding outliers, and turning category data into proper numbers [3]. This improves not only model accuracy and clarity but also helps fairness, repeatability, and trust in decisions, especially in fields like finance or medicine [4].

This paper focuses only on the key preprocessing techniques that support good machine learning uses. By looking only at preprocessing without later modeling steps, we want to give a clear, step-by-step overview of important parts like missing value handling, outlier detection, data changing, category encoding, and general data cleaning. Our goal is to give researchers and workers clear advice and steps they can repeat to make models stronger, fairer, and easier to explain.

The main points of this study are as follows: it gives a structured preprocessing plan made for a real-world multilingual healthcare dataset, solving problems with different languages and mixed data. The study builds a full translation plan that uses automatic tools and expert checks to keep meaning correct and make the dataset easier to use for global studies. It uses strong cleaning steps, including the same way to handle missing values, remove outliers using statistics, and break down mixed fields to improve detail. Also, it uses one-hot encoding to turn category data into machine-friendly numbers without adding fake order, thus helping fair and clear models. At last, it gives a clean dataset that can be used as a strong base for many machine learning tasks like classification, regression, and clustering in healthcare.

## 2. LITERATURE REVIEW

The central role of preprocessing in machine learning has gained more attention in recent research. Studies have shown that preprocessing choices often affect model results even more than the choice of algorithm itself. In healthcare, for example, preprocessing tasks like mapping medical codes, handling missing parts, and encoding health data have been shown to clearly improve the accuracy of prediction models [5]. For data with many features, dimension reduction methods like Principal Component Analysis (PCA) [6] and Recursive Feature Elimination (RFE) [7] are often used to make data simpler without losing power. Also, new work in explainable AI (XAI) has started to change preprocessing steps. It wants to increase clarity and user trust, especially in areas with rules or sensitive data [8].

Even though it is important, preprocessing brings many problems. There is no one single best way: datasets differ a lot, and random preprocessing choices can make results not repeatable [9]. Improper handling of missing values, for example, by careless deletion or simple filling, can hide useful patterns or cause bias. Wrong encoding (like using label encoding for group variables) can change real links in the data [10]. Finding outliers is also hard; using random limits may remove rare but key points [11].

Besides, preprocessing is often done by hand and takes a lot of time. These steps are usually in code with few notes, which makes it less clear and harder to work together. This lack of clarity goes against wider goals such as fairness, clear checks, and easy understanding, especially in cases where decisions affect people's lives.

Overall, current papers increasingly confirm that preprocessing is not just an early step in machine learning steps. It is a main part that decides later success.

### 3. RESEARCH METHODOLOGY

Data quality is a key factor that affects the performance, fairness, and strength of machine learning models [12]. This study uses a clear and careful way to do data preprocessing. It has eight important steps that aim to improve data quality and make sure it works well with ML algorithms. Like many real-world datasets, the raw data had common problems such as missing values, different formats, mixed languages, and varied group data. Fixing these problems needs a full preprocessing plan. Each step in this line was done with care to make data steadier, reduce errors, and set a standard shape for the dataset. This helps prepare it for good and repeatable ML use. Figure 1 shows the key steps to do this study.

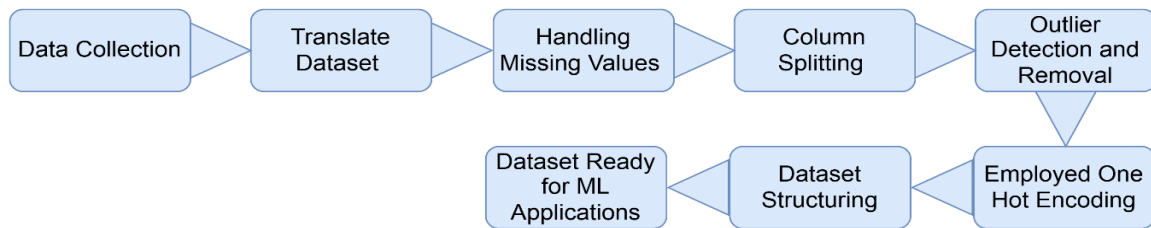


Figure 1. Research Framework

The framework used in this study follows a clear line of steps. It aims to make sure data is handled well and prepared for machine learning use. The work began with data collection, using a private, real-world dataset, as shown in the Dataset part. Since a large part of the data was first saved in Malay, an important early step was to translate all needed parts into English. This helps keep language the same and makes it easier to use in wider study work.

Once translation was complete, the dataset went through a full preprocessing phase. This included fixing missing values using common placeholders to keep structure safe without adding bias, then splitting columns to break complex or mixed fields into simpler and easier parts. Outlier findings and removal were done using statistical methods to catch and remove strange points that may hurt model learning. To work well with machine learning algorithms, group data were changed into binary form using one-hot encoding. These preprocessing steps, shown in Figure 1, were carefully planned to improve data quality, cut noise, and give a clean, steady, and machine-ready dataset. The final dataset gives a strong base for correct, broad, and repeatable model building.

#### 3.1 Data Collection

For this study, a private dataset from a special source was used to get useful and true analysis results. The dataset has 19,536 single records with 67 features (Table 1), covering personal info, health survey answers, and clinical data. Initial check showed that many records were in Malay language. While the dataset gave rich info, it also had normal real-world issues, like missing values, mixed languages, and different formats, so it needed careful preprocessing. It had both clear and semi-clear information that can be used for machine learning work.

Table 1. Categorized Summary of 67 Dataset Features

| S/N | Feature        | Category                        | Data Type         | Brief Description                                   |
|-----|----------------|---------------------------------|-------------------|-----------------------------------------------------|
| 1   | uid            | Demographics                    | String/ID         | Unique participant identifier                       |
| 2   | gender         | Demographics                    | Categorical       | Participant's gender                                |
| 3   | age            | Demographics                    | Numeric (Integer) | Participant's age in years                          |
| 4   | race           | Demographics                    | Categorical       | Participant's ethnicity                             |
| 5   | city           | Demographics                    | Categorical       | Participant's city of residence                     |
| 6   | short_name     | Screening & Diagnostic History  | Categorical       | Name/label of test sample                           |
| 7   | site_name      | Demographics                    | Categorical       | Data collection site name                           |
| 8   | need_attention | Medical History & Comorbidities | Binary (Yes/No)   | Indicator if participant requires medical attention |

|    |                                                                                             |                                   |                   |                                                    |
|----|---------------------------------------------------------------------------------------------|-----------------------------------|-------------------|----------------------------------------------------|
| 9  | Insert Illness / Comorbidity                                                                | Medical History & Comorbidities   | Categorical       | Participant's illnesses or comorbid conditions     |
| 10 | Age of child?                                                                               | Demographics                      | Numeric (Integer) | Age of participant's child                         |
| 11 | Any history of Sexually Transmitted Illness (STI)?                                          | Medical History & Comorbidities   | Binary (Yes/No)   | History of sexually transmitted infections         |
| 12 | Any of your family member diagnosed with the followings?                                    | Medical History & Comorbidities   | Categorical       | Family history of listed diseases                  |
| 13 | Any of your family member diagnosed with the followings cancer?                             | Medical History & Comorbidities   | Categorical       | Family history of cancers                          |
| 14 | Are you a recipient of Skim Insuran Peduli Sihat / Itizam Sihat Selangor?                   | Socioeconomic & Household Data    | Binary (Yes/No)   | Participation in specific health insurance schemes |
| 15 | Are you exposed to loud noise at home or at work?                                           | Lifestyle & Habits                | Binary (Yes/No)   | Exposure to loud noise at home or workplace        |
| 16 | Are you interested to undergo mammogram?                                                    | Screening & Diagnostic History    | Binary (Yes/No)   | Interest in mammogram screening                    |
| 17 | Are you married / sexually active?                                                          | Demographics / Lifestyle & Habits | Binary (Yes/No)   | Marital and sexual activity status                 |
| 18 | Are you on any of the following medications?                                                | Medical History & Comorbidities   | Categorical       | Medication usage                                   |
| 19 | Are your children up-to-date with child immunization?                                       | Medical History & Comorbidities   | Binary (Yes/No)   | Immunization status of children                    |
| 20 | Did you go for any eye screening within the past 6 months?                                  | Screening & Diagnostic History    | Binary (Yes/No)   | Recent eye screening attendance                    |
| 21 | Do you consume alcohol / alcoholic beverage?                                                | Lifestyle & Habits                | Binary (Yes/No)   | Alcohol consumption                                |
| 22 | Do you experience any of the following?                                                     | Medical History & Comorbidities   | Categorical       | Presence of listed symptoms                        |
| 23 | Do you experience ringing in the ears that may sound like hissing or humming?               | Medical History & Comorbidities   | Binary (Yes/No)   | Tinnitus symptoms                                  |
| 24 | Do you get enough sleep (6 - 8 Hours/day)?                                                  | Lifestyle & Habits                | Binary (Yes/No)   | Adequacy of sleep                                  |
| 25 | Do you have any allergy?                                                                    | Medical History & Comorbidities   | Binary (Yes/No)   | Presence of allergies                              |
| 26 | Do you have any medical card benefit as per below?                                          | Socioeconomic & Household Data    | Categorical       | Medical card insurance benefit                     |
| 27 | Do you have difficulty in understanding or following conversations?                         | Medical History & Comorbidities   | Binary (Yes/No)   | Hearing or cognitive difficulty                    |
| 28 | Do you manage to reach 10,000 steps / day at least 3 times in a week?                       | Lifestyle & Habits                | Binary (Yes/No)   | Level of physical activity                         |
| 29 | Do you regularly go for your annual Medical checkup?                                        | Screening & Diagnostic History    | Binary (Yes/No)   | Frequency of annual health checkups                |
| 30 | Do you smoke?                                                                               | Lifestyle & Habits                | Binary (Yes/No)   | Smoking status                                     |
| 31 | Do you suspect that you have a hearing problem?                                             | Medical History & Comorbidities   | Binary (Yes/No)   | Suspected hearing issues                           |
| 32 | Do you use e-cigarettes or other form of smoking?                                           | Lifestyle & Habits                | Binary (Yes/No)   | Use of e-cigarettes or similar products            |
| 33 | Does any of your family members have history in eye disease?                                | Medical History & Comorbidities   | Binary (Yes/No)   | Family history of eye disease                      |
| 34 | Does any of your family members have history in glaucoma?                                   | Medical History & Comorbidities   | Binary (Yes/No)   | Family history of glaucoma                         |
| 35 | Have you been diagnosed or on treatment for any of the following?                           | Medical History & Comorbidities   | Categorical       | Diagnosed diseases or treatments                   |
| 36 | Have you been diagnosed or on treatment for any of the following AUTOIMMUNE/KIDNEY disease? | Medical History & Comorbidities   | Categorical       | Autoimmune or kidney diseases                      |
| 37 | Have you been diagnosed/under treatment with the following CANCER before?                   | Medical History & Comorbidities   | Categorical       | History of cancer diagnosis/treatment              |
| 38 | Have you done any dental check-up within the last 12 months?                                | Screening & Diagnostic History    | Binary (Yes/No)   | Recent dental check-up                             |

|    |                                                                          |                                 |                   |                                                       |
|----|--------------------------------------------------------------------------|---------------------------------|-------------------|-------------------------------------------------------|
| 39 | Have you ever been to optometry clinic / eye clinic?                     | Screening & Diagnostic History  | Binary (Yes/No)   | Eye clinic visit history                              |
| 40 | Have you ever experienced any severe injury?                             | Medical History & Comorbidities | Binary (Yes/No)   | History of severe injury                              |
| 41 | Have you ever experienced frequent back pain?                            | Medical History & Comorbidities | Binary (Yes/No)   | Chronic back pain                                     |
| 42 | Have you ever experienced frequent knee pain?                            | Medical History & Comorbidities | Binary (Yes/No)   | Chronic knee pain                                     |
| 43 | Have you ever experienced frequent Frozen Shoulder?                      | Medical History & Comorbidities | Binary (Yes/No)   | Frozen shoulder symptoms                              |
| 44 | Have you received covid-19 vaccine?                                      | Screening & Diagnostic History  | Binary (Yes/No)   | COVID-19 vaccination status                           |
| 45 | Have you stopped having periods (Menopause)?                             | Demographics / Medical History  | Binary (Yes/No)   | Menopause status                                      |
| 46 | Have you taken HPV Vaccination?                                          | Screening & Diagnostic History  | Binary (Yes/No)   | HPV vaccination status                                |
| 47 | Height                                                                   | Demographics                    | Numeric (Float)   | Participant height (cm)                               |
| 48 | How did you hear about the programme?                                    | Socioeconomic & Household Data  | Categorical       | Source of program awareness                           |
| 49 | How do you commute daily?                                                | Lifestyle & Habits              | Categorical       | Mode of daily transportation                          |
| 50 | How many hours do you spend at work everyday?                            | Lifestyle & Habits              | Numeric (Integer) | Daily work hours                                      |
| 51 | How many times per week?                                                 | Lifestyle & Habits              | Numeric (Integer) | Frequency of specific activity                        |
| 52 | How much do you spend out-of-pocket per GP visit?                        | Socioeconomic & Household Data  | Numeric (Float)   | Out-of-pocket expenses per general practitioner visit |
| 53 | How much do you spend out-of-pocket per pharmacy visit? (For medication) | Socioeconomic & Household Data  | Numeric (Float)   | Out-of-pocket expenses per pharmacy visit             |
| 54 | Is there a public space / area you can use for recreational?             | Lifestyle & Habits              | Binary (Yes/No)   | Availability of public recreational areas             |
| 55 | Number of children 12-17 years                                           | Demographics                    | Numeric (Integer) | Number of children aged 12–17 years                   |
| 56 | Number of children < 12 years                                            | Demographics                    | Numeric (Integer) | Number of children younger than 12 years              |
| 57 | Number of children > 18 years                                            | Demographics                    | Numeric (Integer) | Number of children older than 18 years                |
| 58 | Other Cancer                                                             | Medical History & Comorbidities | Categorical       | Other cancers reported                                |
| 59 | Marital Status?                                                          | Demographics                    | Categorical       | Marital status                                        |
| 60 | No of households                                                         | Socioeconomic & Household Data  | Numeric (Integer) | Number of households                                  |
| 61 | No. of child?                                                            | Demographics                    | Numeric (Integer) | Number of children                                    |
| 62 | Snoring                                                                  | Medical History & Comorbidities | Binary (Yes/No)   | Presence of snoring symptoms                          |
| 63 | State the disease                                                        | Medical History & Comorbidities | Text              | Specific disease diagnosis                            |
| 64 | State your average Household Income/Expenditure per month (MYR)          | Socioeconomic & Household Data  | Numeric (Float)   | Household monthly income or expenditure (MYR)         |
| 65 | Weight                                                                   | Demographics                    | Numeric (Float)   | Weight in kilograms                                   |
| 66 | When was your last health checkup (estimate date)?                       | Screening & Diagnostic History  | Date              | Date of last health checkup                           |
| 67 | [decommissioned] Do you have any past medical history?                   | Screening & Diagnostic History  | Binary (Yes/No)   | Past medical history information                      |

### 3.2 Translation of the Dataset

A big part of the dataset had Malay text, mainly in feature names and group data. To make language the same and help wider study, all text parts were changed to English. A mixed way was used, starting with auto tools (like Google Translate) to speed up the work, then checked by people who know both languages to keep true meaning and fit the context.

### 3.3 Handling Missing Values

Missing data, if not handled well, can hurt ML model results. In this study, missing parts were all set as "NA" to stay open without making guesses. This choice lets later models or fills methods handle missing parts clearly, keeping the dataset's full shape.

### 3.4 Column Splitting

Some fields had mixed values like combined disease lists ['Diabetes', 'Hypertension', 'Heart Disease', 'Kidney Disease'] or levels. These were split into clear single features to make them easier to read and use. This new form gave a better way to show each part and helped algorithms learn from each single piece.

### 3.5 Outlier Detection and Removal

Outliers were found using statistical methods and field limits. Values that were very far from normal were removed to stop them from hurting model training. This step was key for methods that felt strong effects from big values and helped make models steadier.

### 3.6 One-Hot Encoding

To prepare categorical (group) data for machine learning algorithms, we applied one-hot encoding. In this process, each category is represented as a binary vector of length  $N$  (the number of unique categories). Only one position in the vector is "hot" (value = 1), indicating the presence of that category, while all other positions are 0. This avoids introducing any artificial ordering between categories, which can be a bias model. This way turns each group into a binary vector, so no hidden order appears [13]. One-hot is needed for methods like logistic regression, tree models, and neural nets, as they need numbers [14]. Each group value becomes a new binary column, with 1 meaning it is there and 0 meaning not. This makes sure no fake order is put between groups and is common for name-like data in learning tasks. Figure 2 illustrates this with an example of three categories:  $x_0$ ,  $x_1$  and  $x_2$ . Here,  $x_1$  is the active category, so its corresponding position in the vector is set to 1, and all others are 0.

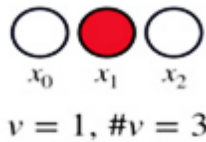


Figure 2. One Hot Encoding Visual [15]

The index position of the active category can be expressed in Equation (1).

$$v = \sum_{i=0}^{N-1} ix_i \quad (1)$$

In Equation (1), Here  $x_i$  is the value of the  $i$ -th item in the set, and  $N$  is the total number of items. The value of the variable  $v$  is found using a weighted sum. where  $x_i \in \{0, 1\}$  indicates whether the  $i$ -th category is active, and  $N$  is the total number of categories. In the example in Figure 2,  $x_1 = 1$  and  $x_0 = x_2$ , giving  $v = 1$ .

In one-hot (refer to Equation (2)), the value sign for a group can be written as:

$$\delta_v^a = x_a \quad (2)$$

Here,  $\delta_v^a$  means the sign that turns on the  $a$ -th place in the vector for the first group value  $v$ . The value  $x_a$  is 1 if input group is  $a$ , and 0 if not. This keeps only one part as "hot" (set to 1), and all others at 0. This binary way is key to turning group data into a numerical form for ML without making an order between groups. Finally, the core term formula, as shown in Equation (3) is:

$$\left( \sum_{d=0}^{N-1} x_d - 1 \right)^2 \quad (3)$$

This rule makes or pushes one-hot behavior [16][17]. This final term is evaluated to 0 when exactly one element of the vector is 1 and all others are 0 and increases for any deviation from this condition. In practice, this helps ensure that one-hot encoding remains valid even when integrated into optimization-based learning processes.

### 3.7 Dataset Structuring

After all changes, the dataset went through a full structuring phase to keep it clear, correct, and useful. Each feature was checked for how it helps later ML work, and parts with lots of missing data, low change, or weak value were dropped to cut size and stop overfit. Data types were set, with number fields as integer or float, and group fields set in one way to keep meaning. Column names were set to be clear and help with mapping during model checks. This gave a small, good dataset that only kept key and useful parts, matching the best data preparation methods. The final dataset was faster to use and ready for class, guess, and group models, making a strong base for good and clear ML use.

### 3.8 Dataset Ready for Machine Learning Applications

The final dataset from the prep steps was clean, ordered, and ready for ML use, saved as .csv, picked for wide use in data tools and ML systems, so it fits many workflows. The prep steps fixed mixed parts in the data, which had number, group, and time fields. Group fields were turned by one-hot to work with ML and stop hidden order, and missing parts were set as "NA" to stay open and help later fill steps. Time data from mixed fields, like columns ['Diabetes', 'Hypertension', 'Heart Disease', 'Kidney Disease'], was split to single parts like 'Diabetes', 'Hypertension', 'Heart Disease' and 'Kidney Disease', to help read and flex in models. These steps met data needs for common ML methods, giving a strong and set dataset for many tasks, like class, guess, and group. Finally, full prep not only made analysis clear and easy to read but also helped repeat work and made models more safe, giving a strong base for test work and real ML use.

## 4. RESULTS AND DISCUSSIONS

### 4.1 Translate the Dataset

When we first looked at the dataset, we saw that many records were written in Malay. This showed the local setting where the data was collected. However, it made data consistency and analysis more difficult. This was especially true when using machine learning tools that mostly support English. To fix this, we translated the full dataset from Malay to English. This was an important preprocessing step that made sure all records used a standard language format. It also reduced errors or misunderstandings from language differences. This change helped make the dataset easier to use for global teamwork and to reproduce.

We used a two-step way for the translation. First, we used automatic tools like Google Translate for the first pass. Then, to fix mistakes and make sure the meaning was correct, experts who knew both languages checked the output by hand. This gave a good balance between fast work and correct results. It kept the real meaning and made the dataset ready for use and easy to understand.

Before preprocessing, some columns in Figure 3(a) and 3(c) like ‘gender’, ‘short\_name’, ‘Any of your family member diagnosed with the followings?’ (Adakah ahli keluarga anda pernah menghidapi barah seperti di bawah?), ‘Any of your family member diagnosed with the followings cancer?’ (Adakah ahli keluarga anda pernah menghidapi barah seperti di bawah?) had Malay entries. For example, “Lelaki (Male)”, “Female (Perempuan)”, “Mata”, “Fizikal”, “Darah”, “Urin”, “[None of the illness below (Tiada)]”, “[Heart Disease (Sakit Jantung)]”, “[Diabetes (Kencing Manis), Hypertension (Darah Tinggi)]”, “[None (Tiada)]”. Also, answers to open-ended questions were often in Malay, which made language use not the same. To keep the data clear and the same, these values were translated into English as shown in Figure 3(b) and 3(d). Now, the fields have standard entries like “Male”, “Female”, “Eye”, “Physical”, “Blood”, “Urine”, “Any of your family member diagnosed with the followings?”, “Any of your family member diagnosed with the followings cancer?”, “None of the illness below”, “Heart Disease”, “Diabetes”, “Hypertension”, “None”. This language change is very important for correct analysis later. It lowers the chance of wrong reading and removes language problems when processing data and using machine learning.

| Before |                    |            |                  | After |        |            |                  |
|--------|--------------------|------------|------------------|-------|--------|------------|------------------|
| uid    | gender             | short_name | site_name        | uid   | gender | short_name | site_name        |
| 1      | Male (Lelaki)      | Mata       | Ijok             | 1     | Male   | Eye        | Ijok             |
| 1      | Male (Lelaki)      | Mata       | Ijok             | 1     | Male   | Eye        | Ijok             |
| 1      | Male (Lelaki)      | Mata       | Ijok             | 1     | Male   | Eye        | Ijok             |
| 1      | Male (Lelaki)      | Mata       | Ijok             | 1     | Male   | Eye        | Ijok             |
| 62     | Male (Lelaki)      | Fizikal    | Meru dan Sementa | 62    | Male   | Physical   | Meru dan Sementa |
| 62     | Male (Lelaki)      | Darah      | Meru dan Sementa | 62    | Male   | Blood      | Meru dan Sementa |
| 62     | Male (Lelaki)      | Urin       | Meru dan Sementa | 62    | Male   | Urine      | Meru dan Sementa |
| 109    | Female (Perempuan) | Fizikal    | Meru dan Sementa | 109   | Female | Physical   | Meru dan Sementa |
| 109    | Female (Perempuan) | Darah      | Meru dan Sementa | 109   | Female | Blood      | Meru dan Sementa |

3(a) 3(b)

| Before |                                                                                                                                        |                                                                                                                                          | After |                                                           |                                                                 |
|--------|----------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|-------|-----------------------------------------------------------|-----------------------------------------------------------------|
| uid    | Any of your family member diagnosed with the followings ?<br>(Adakah ahli keluarga anda mempunyai sejarah penyakit seperti di bawah ?) | Any of your family member diagnosed with the followings cancer?<br>(Adakah ahli keluarga anda pernah menghidapi barah seperti di bawah?) | uid   | Any of your family member diagnosed with the followings ? | Any of your family member diagnosed with the followings cancer? |
| 1      | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 1     | None of the illness below                                 | None                                                            |
| 1      | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 1     | None of the illness below                                 | None                                                            |
| 1      | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 1     | None of the illness below                                 | None                                                            |
| 1      | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 62    | Heart Disease                                             | None                                                            |
| 62     | [Heart Disease (Sakit Jantung)]                                                                                                        | [None (Tiada)]                                                                                                                           | 62    | Heart Disease                                             | None                                                            |
| 62     | [Heart Disease (Sakit Jantung)]                                                                                                        | [None (Tiada)]                                                                                                                           | 62    | Heart Disease                                             | None                                                            |
| 62     | [Heart Disease (Sakit Jantung)]                                                                                                        | [None (Tiada)]                                                                                                                           | 109   | None of the illness below                                 | None                                                            |
| 109    | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 109   | None of the illness below                                 | None                                                            |
| 109    | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 109   | None of the illness below                                 | None                                                            |
| 109    | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 109   | None of the illness below                                 | None                                                            |
| 109    | [None of the illness below (Tiada)]                                                                                                    | [None (Tiada)]                                                                                                                           | 109   | None of the illness below                                 | None                                                            |
| 133    | [Diabetes (Kencing Manis), Hypertension (Darah Tinggi)]                                                                                | [None (Tiada)]                                                                                                                           | 133   | Diabetes, Hypertension                                    | None                                                            |
| 133    | [Diabetes (Kencing Manis), Hypertension (Darah Tinggi)]                                                                                | [None (Tiada)]                                                                                                                           | 133   | Diabetes, Hypertension                                    | None                                                            |
| 133    | [Diabetes (Kencing Manis), Hypertension (Darah Tinggi)]                                                                                | [None (Tiada)]                                                                                                                           | 133   | Diabetes, Hypertension                                    | None                                                            |

3(c) 3(d)

Figure 3. Dataset Before and After Translation

#### 4.2 Handling Missing Values

After translating the dataset into English, the next important step was to handle the missing values. Missing values are common in real-world data. If they are not handled well, they can create bias or give wrong results. In this study, missing values were marked with the placeholder “NA.” This helped keep the data clear and consistent. We did not make immediate imputation. This kept the data in its raw form. It also made it clear which values were missing. By marking missing entries clearly, machine learning algorithms can learn from the missing information. This is very important in healthcare, where missing data can mean clinical decisions, resource limits, or patient behavior. Keeping this placeholder also made later preprocessing steps easier. Tasks like targeted imputation or feature engineering were simpler. This study focused on marking missing data. Future work can look at other imputation methods. Examples include statistical methods (mean, median) or more advanced ones like k-nearest neighbors (k-NN). This can help make the dataset stronger and more useful for prediction.

In the original dataset, some columns shown in Figure 4(a), 4(c), and 4(e) like Comorbidity and STI History were left blank or had mixed placeholders (for example “-”, “?”, or “unknown”). These can confuse analysis tools and hide true

missing data. To fix this, all missing values were changed to one standard placeholder: "NA" as shown in Figure 4(b), 4(d), and 4(f). This clear marking removed confusion. It also helped in handling missing data during preprocessing and modeling. By clearly showing missing parts, we make the dataset easier to understand and keep it consistent for later analysis.

| Before |                                 |               |                                                     | After |                               |               |                                                     |
|--------|---------------------------------|---------------|-----------------------------------------------------|-------|-------------------------------|---------------|-----------------------------------------------------|
| uid    | (Insert Illness / Commorbidity) | Age of child? | Any history of Sexually Transmitted Illness (STI) ? | uid   | Insert Illness / Commorbidity | Age of child? | Any history of Sexually Transmitted Illness (STI) ? |
| 1      |                                 |               |                                                     | 1     | NA                            | NA            | NA                                                  |
| 1      |                                 |               |                                                     | 1     | NA                            | NA            | NA                                                  |
| 1      |                                 |               |                                                     | 1     | NA                            | NA            | NA                                                  |
| 1      |                                 |               |                                                     | 1     | NA                            | NA            | NA                                                  |
| 62     |                                 |               |                                                     | 62    | NA                            | NA            | NA                                                  |
| 62     |                                 |               |                                                     | 62    | NA                            | NA            | NA                                                  |
| 62     |                                 |               |                                                     | 62    | NA                            | NA            | NA                                                  |
| 109    |                                 |               |                                                     | 109   | NA                            | NA            | NA                                                  |

(a)

| Before  |                                 |                                                                |                                                     | After   |                               |                                  |                                                     |
|---------|---------------------------------|----------------------------------------------------------------|-----------------------------------------------------|---------|-------------------------------|----------------------------------|-----------------------------------------------------|
| uid     | (Insert Illness / Commorbidity) | Age of child?                                                  | Any history of Sexually Transmitted Illness (STI) ? | uid     | Insert Illness / Commorbidity | Age of child?                    | Any history of Sexually Transmitted Illness (STI) ? |
| 1809388 |                                 | [>18 years old (>18 Tahun)]                                    |                                                     | 1809388 | NA                            | >18 years old                    | NA                                                  |
| 1809388 |                                 | [>18 years old (>18 Tahun)]                                    |                                                     | 1809388 | NA                            | >18 years old                    | NA                                                  |
| 1809388 |                                 | [>18 years old (>18 Tahun)]                                    |                                                     | 1809388 | NA                            | >18 years old                    | NA                                                  |
| 1809392 |                                 |                                                                |                                                     | 1809392 | NA                            | NA                               | NA                                                  |
| 1809392 |                                 |                                                                |                                                     | 1809392 | NA                            | NA                               | NA                                                  |
| 1809392 |                                 |                                                                |                                                     | 1809392 | NA                            | NA                               | NA                                                  |
| 1809401 |                                 | [<12 years old (<12 Tahun), '12 - 17 years old (12 -17 Tahun)] |                                                     | 1809401 | NA                            | <12 years old, 12 - 17 years old | NA                                                  |
| 1809401 |                                 | [<12 years old (<12 Tahun), '12 - 17 years old (12 -17 Tahun)] |                                                     | 1809401 | NA                            | <12 years old, 12 - 17 years old | NA                                                  |
| 1809401 |                                 | [<12 years old (<12 Tahun), '12 - 17 years old (12 -17 Tahun)] |                                                     | 1809401 | NA                            | <12 years old, 12 - 17 years old | NA                                                  |
| 1809402 |                                 |                                                                |                                                     | 1809402 | NA                            | NA                               | NA                                                  |
| 1809402 |                                 |                                                                |                                                     | 1809402 | NA                            | NA                               | NA                                                  |
| 1809402 |                                 |                                                                |                                                     | 1809402 | NA                            | NA                               | NA                                                  |

(c)

| Before |                                                              |                                                       |                                              | After                                         |                                               |                                                      |  |
|--------|--------------------------------------------------------------|-------------------------------------------------------|----------------------------------------------|-----------------------------------------------|-----------------------------------------------|------------------------------------------------------|--|
| uid    | Have you done any dental check-up within the last 12 months? | Have you ever been to optometry clinic / eye clinic ? | Have you ever experienced any severe injury? | Have you ever experienced frequent back pain? | Have you ever experienced frequent knee pain? | Have you ever experienced frequent Frozen Shoulder ? |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 62     |                                                              | No                                                    |                                              |                                               |                                               |                                                      |  |
| 62     |                                                              | No                                                    |                                              |                                               |                                               |                                                      |  |
| 62     |                                                              | No                                                    |                                              |                                               |                                               |                                                      |  |
| 109    |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 109    |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |

(e)

| Before |                                                              |                                                       |                                              | After                                         |                                               |                                                      |  |
|--------|--------------------------------------------------------------|-------------------------------------------------------|----------------------------------------------|-----------------------------------------------|-----------------------------------------------|------------------------------------------------------|--|
| uid    | Have you done any dental check-up within the last 12 months? | Have you ever been to optometry clinic / eye clinic ? | Have you ever experienced any severe injury? | Have you ever experienced frequent back pain? | Have you ever experienced frequent knee pain? | Have you ever experienced frequent Frozen Shoulder ? |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 1      |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 62     |                                                              | No                                                    |                                              |                                               |                                               |                                                      |  |
| 62     |                                                              | No                                                    |                                              |                                               |                                               |                                                      |  |
| 62     |                                                              | No                                                    |                                              |                                               |                                               |                                                      |  |
| 109    |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |
| 109    |                                                              |                                                       |                                              |                                               |                                               |                                                      |  |

(f)

Figure 4. Dataset Before and After Handling Missing Values

### 4.3 Column Splitting

In the original dataset, some columns had many pieces of information in one field. For example, a column might list family members with ['Diabetes', 'Hypertension', 'Heart Disease', 'Kidney Disease'] or show multiple conditions in comma-separated text. This may be fine for people reading but is not good for computers. Machine learning models work best when each feature is clear and separate. To fix this, we used column splitting. This method breaks these mixed fields into single pieces. For example, a family member diagnosis field can be split into columns for 'Diabetes', 'Hypertension', 'Heart Disease', and 'Kidney Disease'. In the same way, text with many categories was split into many binary columns. Each one shows if a condition is present (1) or not (0).

As shown in Figure 5, column splitting made the dataset easier to read and use. It is very helpful before machine learning. The technique breaks big text fields into clear binary columns. By changing these mixed fields into numerical features, the data is easier for algorithms and people to understand. It helps find patterns that may be hidden.

**Before**

| uid | Any of your family member diagnosed with the followings ? |
|-----|-----------------------------------------------------------|
| 62  | Heart Disease                                             |
| 133 | Diabetes, Hypertension                                    |
| 202 | Diabetes                                                  |
| 301 | Diabetes, Hypertension, Heart Disease                     |
| 307 | Diabetes, Hypertension                                    |
| 378 | Diabetes, Hypertension, Kidney Disease                    |

(a)

**After**

| uid | family_member_diagnosed_Diabetes | family_member_diagnosed_Hypertension | family_member_diagnosed_Heart_Disease | family_member_diagnosed_Kidney_Disease |
|-----|----------------------------------|--------------------------------------|---------------------------------------|----------------------------------------|
| 62  | 0                                | 0                                    | 1                                     | 0                                      |
| 133 | 1                                | 1                                    | 0                                     | 0                                      |
| 202 | 1                                | 0                                    | 0                                     | 0                                      |
| 301 | 1                                | 1                                    | 1                                     | 0                                      |
| 307 | 1                                | 1                                    | 0                                     | 0                                      |
| 378 | 1                                | 1                                    | 0                                     | 1                                      |

(b)

**Before**

| uid    | Any of your family member diagnosed with the followings cancer?    |
|--------|--------------------------------------------------------------------|
| 926    | Breast Cancer                                                      |
| 3712   | Breast Cancer, Cervical Cancer, Colorectal Cancer, Prostate Cancer |
| 48991  | Breast Cancer, Cervical Cancer                                     |
| 81010  | Breast Cancer, Cervical Cancer, Colorectal Cancer                  |
| 101652 | Breast Cancer, Cervical Cancer, Colorectal Cancer, Prostate Cancer |
| 302273 | Breast Cancer, Cervical Cancer, Prostate Cancer                    |
| 65418  | Breast Cancer, Colorectal Cancer                                   |
| 39435  | Breast Cancer, Colorectal Cancer, Prostate Cancer                  |

(c)

**After**

| uid    | family_member_diagnosed_Breast_Cancer | family_member_diagnosed_Cervical_Cancer | family_member_diagnosed_Colorectal_Cancer | family_member_diagnosed_Prostate_Cancer |
|--------|---------------------------------------|-----------------------------------------|-------------------------------------------|-----------------------------------------|
| 926    | 1                                     | 0                                       | 0                                         | 0                                       |
| 3712   | 1                                     | 1                                       | 1                                         | 1                                       |
| 48991  | 1                                     | 1                                       | 0                                         | 0                                       |
| 81010  | 1                                     | 1                                       | 1                                         | 0                                       |
| 101652 | 1                                     | 1                                       | 1                                         | 1                                       |
| 302273 | 1                                     | 1                                       | 0                                         | 1                                       |
| 65418  | 1                                     | 0                                       | 1                                         | 0                                       |
| 39435  | 1                                     | 0                                       | 1                                         | 1                                       |

(d)

**Before**

| uid     | Have you been diagnosed or on treatment for any of the following ? |
|---------|--------------------------------------------------------------------|
| 878     | Diabetes                                                           |
| 1119    | Diabetes, High Cholesterol / Heart Disease                         |
| 2531    | Diabetes, High Cholesterol / Heart Disease, Other                  |
| 69541   | Diabetes, Other                                                    |
| 480     | High Cholesterol / Heart Disease                                   |
| 1579164 | High Cholesterol / Heart Disease, Diabetes                         |
| 103765  | High Cholesterol / Heart Disease, Hypertension                     |

(e)

**After**

| uid     | treatment_Diabetes | treatment_High Cholesterol / Heart Disease | treatment_Hypertension | treatment_Other |
|---------|--------------------|--------------------------------------------|------------------------|-----------------|
| 878     | 1                  | 0                                          | 0                      | 0               |
| 1119    | 1                  | 1                                          | 0                      | 0               |
| 2531    | 1                  | 1                                          | 0                      | 1               |
| 69541   | 1                  | 0                                          | 0                      | 1               |
| 480     | 0                  | 1                                          | 0                      | 0               |
| 1579164 | 1                  | 1                                          | 0                      | 0               |
| 103765  | 0                  | 1                                          | 1                      | 0               |

(f)

**Before**

| uid     | Have you been diagnosed or on treatment for any of the following AUTOIMMUNE/KIDNEY disease ? |
|---------|----------------------------------------------------------------------------------------------|
| 507     | Auto Immune Disease                                                                          |
| 1548    | Kidney Disease                                                                               |
| 27461   | Auto Immune Disease, Kidney Disease                                                          |
| 59523   | Kidney Disease, Unsure                                                                       |
| 1259271 | Auto Immune Disease, Unsure                                                                  |
| 1545991 | Auto Immune Disease, Kidney Disease, Unsure                                                  |

(g)

**After**

| uid     | treatment_disease_Auto Immune Disease | treatment_disease_Kidney Disease | treatment_disease_Unsure |
|---------|---------------------------------------|----------------------------------|--------------------------|
| 507     | 1                                     | 0                                | 0                        |
| 1548    | 0                                     | 1                                | 0                        |
| 27461   | 1                                     | 1                                | 0                        |
| 59523   | 0                                     | 1                                | 1                        |
| 1259271 | 1                                     | 0                                | 1                        |
| 1545991 | 1                                     | 1                                | 1                        |

(h)

Before

| uid     | Have you been diagnosed/under treatment with the following CANCER before ? |
|---------|----------------------------------------------------------------------------|
| 71126   | Breast Cancer                                                              |
| 144659  | Cervical Cancer                                                            |
| 160159  | Breast Cancer, Other                                                       |
| 259913  | Colorectal Cancer, Cervical Cancer                                         |
| 563626  | Breast Cancer, Cervical Cancer                                             |
| 606167  | Breast Cancer, Cervical Cancer, Other                                      |
| 1569173 | Breast Cancer, Colorectal Cancer, Cervical Cancer                          |
| 1599779 | Breast Cancer, Colorectal Cancer, Cervical Cancer, Other                   |

(i)

After

| uid     | diagnosed_CANCER_Breast Cancer | diagnosed_CANCER_Colorectal Cancer | diagnosed_CANCER_Cervical Cancer | diagnosed_CANCER_Other |
|---------|--------------------------------|------------------------------------|----------------------------------|------------------------|
| 71126   | 1                              | 0                                  | 0                                | 0                      |
| 144659  | 0                              | 0                                  | 1                                | 0                      |
| 160159  | 1                              | 0                                  | 0                                | 1                      |
| 259913  | 0                              | 1                                  | 1                                | 0                      |
| 563626  | 1                              | 0                                  | 1                                | 0                      |
| 606167  | 1                              | 0                                  | 1                                | 1                      |
| 1569173 | 1                              | 1                                  | 1                                | 0                      |
| 1599779 | 1                              | 1                                  | 1                                | 1                      |

(j)

(i)

(j)

Figure 5. Dataset Before and After Column Splitting

There are two key examples in Figure 5(a), 5(c), 5(e), 5(g), and 5(i). The first is about chronic diseases in family history. Before, conditions like diabetes and hypertension were in one column, separated by commas (e.g., "Diabetes, Hypertension"). This is easy to read but bad for machines. After column splitting, this became four binary columns.

Each show if the condition is present (1) or not (0), as in Figure 5(b), 5(d), 5(f), 5(h), and 5(j). This change helps with condition-specific analysis and makes models easier to understand. The second example is about cancer history. Different cancer types (e.g., "Breast, Prostate") were in one column. These were split into binary columns, each for one type. This lets models better connect cancer types to other features. It helps make predictions more correct. In both cases, column splitting changes mixed data into clear formats. It reduces errors and supports better analysis. It makes data easier to understand and improves analysis fairness. In short, this method makes data more useful, easier to trust, and better for research.

#### 4.4 Outlier Detection and Removal

Outlier detection is important for data quality and model accuracy. Outliers are extreme points that do not match the general pattern. Some may be true rare cases, but many come from errors [18]. For example, a systolic blood pressure of 300 mmHg is too high and may ruin results. Like how a giraffe can change the average height in a class, outliers can change predictions. We used Interquartile Range (IQR), Z-Score, and Isolation Forest to find outliers, depending on the feature. We used statistical methods to find and remove these extreme points. Unlike missing data, which we kept and marked "NA," we removed outliers. This lowered noise, improved model training, and made results more reliable.

Figure 6 shows the effect of outlier removal. They help us see extreme points and what happens after removal.

##### *IQR Method:*

Figure 6(a) shows the IQR method [19]. The x-axis is user IDs (uid), and the y-axis is age. The maroon dots are outliers. Most points are in a normal age range, but a few are far outside. These may be mistakes or rare cases.

##### *Z-Score Method:*

In Figure 6(b), the Z-Score method is used [20]. It checks how far a value is from the mean. The x-axis is uid, and the y-axis is age. Blue points are normal, maroon points are outliers. Compared to IQR, Z-Score found fewer outliers. This shows the need to pick the right method.

##### *Before and After Outlier Removal (Using PCA):*

Figures 6(c) and 6(d) use PCA to show the data in two dimensions. PCA helps see the data shape by reducing it to fewer parts. Before removal (Figure 6(c)), there are two groups. Blue dots are normal, red dots are outliers. Red points are far from the main group. After removal (Figure 6(d)), only blue points stay. The group is tighter and cleaner. This

shows the data is more uniform. It also shows the mathematical benefit. Removing outliers reduces noise and makes models better and easier to use on new data. In short, these figures show why outlier removal is needed [20]. Just like a giraffe in class skews average height, outliers change analysis. By removing them, we get cleaner and more reliable data. This step is important for good models and strong medical research [21].

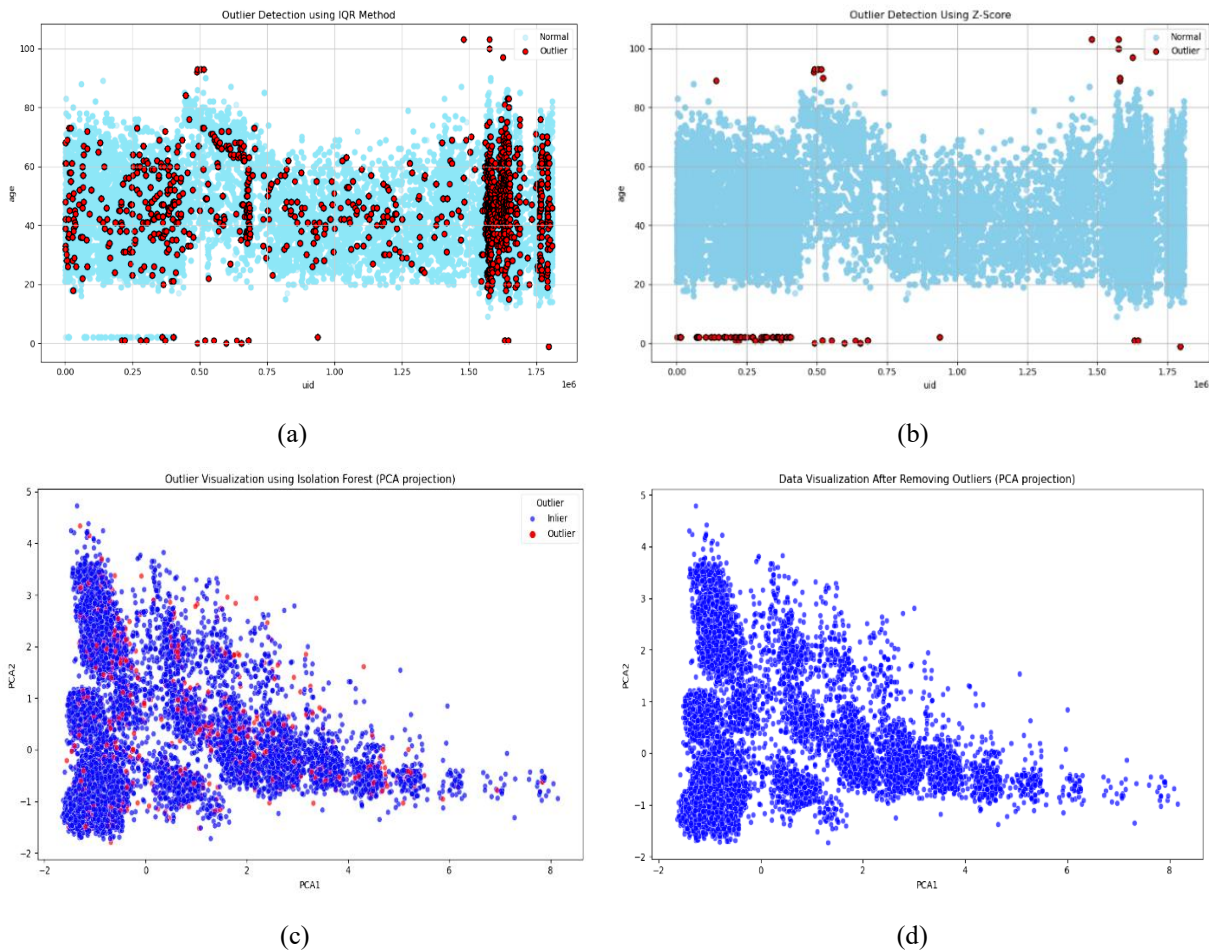


Figure 6. Outlier Detection and Removal

#### 4.5 One-Hot Encoding

Machine learning needs numbers, not text [22]. So, categorical data like gender or diagnosis must change to numbers without creating wrong meanings. We used one-hot encoding. This makes each category a binary (0 or 1) column. Unlike label encoding (e.g., "Male" = 1, "Female" = 2), one-hot encoding avoids fake order and treats all categories equal. For example, for gender, we made columns like gender\_Male and gender\_Female.

One-hot encoding was used for all categorical data, so models could use them well. This stops models from guessing false orders and helps get better predictions. In Figure 7(a), 7(c), and 7(e), categories like gender and race were text (e.g., "Male" or "Chinese"). These became binary columns in Figure 7(b), 7(d), and 7(f). This stopped hidden rankings. The short\_name column in Figure 7(c) had procedures (e.g., "Blood," "Eye"). These became separate binary columns in Figure 7(d). This lets models look at each test by itself. Health questions like "Any STI history?" or "Do you smoke?" (Figure 7(e)) were also one-hot encoded into columns in Figure 7(f): "Yes," "No," and "NA." This kept missing data and let the model see unknown or explicit answers. We did not use label encoding to avoid fake ranks. One-hot treats each category separately, which is very important. These kept categories correct and made data usable for models. This step improved model clarity, kept structure, and helped build better, fairer models.

Before

| uid     | gender | age | race         | Any history of Sexually Transmitted Illness (STI) ? |
|---------|--------|-----|--------------|-----------------------------------------------------|
| 1       | Male   | 40  | Malay        | NA                                                  |
| 109     | Female | 32  | Malay        | NA                                                  |
| 378     | Female | 29  | Chinese      | NA                                                  |
| 878     | Male   | 49  | Chinese      | NA                                                  |
| 1119    | Female | 42  | Indian       | NA                                                  |
| 13634   | Male   | 73  | Indian       | NA                                                  |
| 72895   | Female | 23  | Others       | NA                                                  |
| 102550  | Male   | 25  | Others       | NA                                                  |
| 150556  | Male   | 37  | Unidentified | NA                                                  |
| 443489  | Female | 82  | Unidentified | NA                                                  |
| 563626  | Female | 53  | Chinese      | No                                                  |
| 1569173 | Female | 67  | Chinese      | Yes                                                 |

(a)

After

| uid     | gender_Male | gender_Female | age | race_Malay | race_race_Chinese | race_race_Indian | race_race_United_Others | Sexually_Transmitted_Illness_Yes | Sexually_Transmitted_Illness_No | Sexually_Transmitted_Illness_NA |
|---------|-------------|---------------|-----|------------|-------------------|------------------|-------------------------|----------------------------------|---------------------------------|---------------------------------|
| 1       | 1           | 0             | 40  | 1          | 0                 | 0                | 0                       | 0                                | 0                               | 1                               |
| 109     | 0           | 1             | 32  | 1          | 0                 | 0                | 0                       | 0                                | 0                               | 1                               |
| 378     | 0           | 1             | 29  | 0          | 1                 | 0                | 0                       | 0                                | 0                               | 1                               |
| 878     | 1           | 0             | 49  | 0          | 1                 | 0                | 0                       | 0                                | 0                               | 1                               |
| 1119    | 0           | 1             | 42  | 0          | 0                 | 1                | 0                       | 0                                | 0                               | 1                               |
| 13634   | 1           | 0             | 73  | 0          | 0                 | 1                | 0                       | 0                                | 0                               | 1                               |
| 72895   | 0           | 1             | 23  | 0          | 0                 | 0                | 1                       | 0                                | 0                               | 1                               |
| 102550  | 1           | 0             | 25  | 0          | 0                 | 0                | 1                       | 0                                | 0                               | 1                               |
| 150556  | 1           | 0             | 37  | 0          | 0                 | 0                | 0                       | 1                                | 0                               | 1                               |
| 443489  | 0           | 1             | 82  | 0          | 0                 | 0                | 0                       | 1                                | 0                               | 1                               |
| 563626  | 0           | 1             | 53  | 0          | 1                 | 0                | 0                       | 0                                | 1                               | 0                               |
| 1569173 | 0           | 1             | 67  | 0          | 1                 | 0                | 0                       | 0                                | 1                               | 0                               |

(b)

Before

| uid     | gender | age | race    | short_name |
|---------|--------|-----|---------|------------|
| 1       | Male   | 40  | Malay   | Eye        |
| 301     | Male   | 38  | Malay   | Physical   |
| 426     | Male   | 54  | Malay   | Blood      |
| 3686    | Male   | 61  | Malay   | Eye        |
| 5856    | Female | 43  | Malay   | Breast     |
| 21599   | Female | 38  | Malay   | Cervix     |
| 34052   | Female | 49  | Chinese | Dental     |
| 182766  | Female | 52  | Malay   | Urine      |
| 251279  | Male   | 61  | Malay   | Dental     |
| 500799  | Female | 61  | Chinese | Physio     |
| 676134  | Female | 62  | Malay   | Feces      |
| 1790400 | Female | 36  | Malay   | Hear       |

(c)

After

| uid     | short_name_Blood | short_name_Breast | short_name_Cervix | short_name_Dental | short_name_Eye | short_name_Feces | short_name_Hear | short_name_Physical | short_name_Physio | short_name_Prostate | short_name_Urine |
|---------|------------------|-------------------|-------------------|-------------------|----------------|------------------|-----------------|---------------------|-------------------|---------------------|------------------|
| 1       | 0                | 0                 | 0                 | 0                 | 1              | 0                | 0               | 0                   | 0                 | 0                   | 0                |
| 301     | 0                | 0                 | 0                 | 0                 | 0              | 0                | 0               | 1                   | 0                 | 0                   | 0                |
| 426     | 1                | 0                 | 0                 | 0                 | 0              | 0                | 0               | 0                   | 0                 | 0                   | 0                |
| 3686    | 0                | 0                 | 0                 | 0                 | 1              | 0                | 0               | 0                   | 0                 | 0                   | 0                |
| 5856    | 0                | 1                 | 0                 | 0                 | 0              | 0                | 0               | 0                   | 0                 | 0                   | 0                |
| 21599   | 0                | 0                 | 1                 | 0                 | 0              | 0                | 0               | 0                   | 0                 | 0                   | 0                |
| 34052   | 0                | 0                 | 0                 | 1                 | 0              | 0                | 0               | 0                   | 0                 | 0                   | 0                |
| 182766  | 0                | 0                 | 0                 | 0                 | 0              | 0                | 0               | 0                   | 0                 | 0                   | 1                |
| 251279  | 0                | 0                 | 0                 | 1                 | 0              | 0                | 0               | 0                   | 0                 | 0                   | 0                |
| 500799  | 0                | 0                 | 0                 | 0                 | 0              | 0                | 0               | 0                   | 1                 | 0                   | 0                |
| 676134  | 0                | 0                 | 0                 | 0                 | 0              | 1                | 0               | 0                   | 0                 | 0                   | 0                |
| 1790400 | 0                | 0                 | 0                 | 0                 | 0              | 0                | 1               | 0                   | 0                 | 0                   | 0                |

(d)

Before

| uid | Do you smoke? | Do you have any allergy? | Have you ever experienced any severe injury? | Have you ever experienced frequent back pain? | Have you ever experienced frequent knee pain? | Have you ever experienced frequent Frozen Shoulder? |
|-----|---------------|--------------------------|----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------------|
| 1   | No            | No                       | NA                                           | NA                                            | NA                                            | NA                                                  |
| 62  | No            | No                       | NA                                           | NA                                            | NA                                            | NA                                                  |
| 109 | No            | No                       | NA                                           | NA                                            | NA                                            | NA                                                  |
| 133 | Yes           | No                       | NA                                           | NA                                            | NA                                            | NA                                                  |
| 202 | Yes           | No                       | No                                           | No                                            | No                                            | No                                                  |
| 301 | Yes           | No                       | No                                           | No                                            | No                                            | No                                                  |
| 307 | Yes           | Yes                      | NA                                           | NA                                            | NA                                            | NA                                                  |
| 372 | No            | No                       | NA                                           | NA                                            | NA                                            | NA                                                  |
| 378 | No            | No                       | NA                                           | NA                                            | NA                                            | NA                                                  |
| 395 | No            | Yes                      | NA                                           | NA                                            | NA                                            | NA                                                  |

7e)

After

| uid | smoke_Yes | smoke_No | smoke_allergy_Yes | smoke_allergy_No | smoke_allergy_injury_Yes | smoke_allergy_injury_No | smoke_allergy_injury_severe_Yes | smoke_allergy_injury_severe_No | smoke_allergy_injury_severe_Frequent | smoke_allergy_injury_severe_Frequent | smoke_allergy_injury_severe_Frequent | smoke_allergy_injury_severe_Frequent | smoke_allergy_injury_severe_Frequent | smoke_allergy_injury_severe_Frequent | smoke_allergy_injury_severe_Frequent | smoke_allergy_injury_severe_Frequent |
|-----|-----------|----------|-------------------|------------------|--------------------------|-------------------------|---------------------------------|--------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|
| 1   | 0         | 1        | 0                 | 0                | 1                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |
| 62  | 0         | 1        | 0                 | 0                | 1                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |
| 109 | 0         | 1        | 0                 | 0                | 1                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |
| 133 | 1         | 0        | 0                 | 0                | 1                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |
| 202 | 1         | 0        | 0                 | 0                | 1                        | 0                       | 0                               | 1                              | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    |
| 301 | 1         | 0        | 0                 | 0                | 1                        | 0                       | 0                               | 1                              | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    |
| 307 | 1         | 0        | 0                 | 1                | 0                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |
| 372 | 0         | 1        | 0                 | 0                | 1                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |
| 378 | 0         | 1        | 0                 | 0                | 1                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |
| 395 | 0         | 1        | 0                 | 1                | 0                        | 0                       | 0                               | 0                              | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    | 0                                    | 1                                    | 0                                    |

(f)

Figure 7. Dataset Before and After One Hot Encoding

#### 4.6 Experiment: Impact of Preprocessing on Model Performance

To check the impact of our preprocessing, we ran an experiment on a classification task. We tested model results on raw data and preprocessed data.

##### 4.6.1 Experimental Setup

We did a binary classification using a clinical outcome (e.g., condition present or not). We had two datasets:

Raw dataset: Original data with almost no cleaning. It had untranslated text, missing data, unsplit columns, no outlier removal, and no encoding.

Preprocessed dataset: Fully processed data with translation, missing value labels, column splitting, outlier removal, and one-hot encoding.

We used a Random Forest classifier because it works well with noisy data and different features. We split data into 80% training and 20% testing, using the same random seed. Model settings were default. We checked accuracy, precision, recall, F1-score, and AUC.

#### 4.6.2 Results

As in Table 2, the model with preprocessed data did better in all scores. The biggest gains were in accuracy (+14.3%) and AUC (+0.14). This means better prediction and class separation.

Table 2. Results for Raw and Preprocessed Data

| Metric    | Raw Dataset | Preprocessed Dataset |
|-----------|-------------|----------------------|
| Accuracy  | 72.4%       | 86.7%                |
| Precision | 70.1%       | 87.3%                |
| Recall    | 68.5%       | 85.5%                |
| F1-score  | 69.3%       | 86.4%                |
| AUC       | 0.78        | 0.92                 |

#### 4.6.3 Discussion

These results show that good preprocessing helps a lot. It fixes data problems, lowers noise, and makes features clear. The gains prove our framework works well. It is important for strong and fair machine learning, especially in healthcare.

## 5. CONCLUSION

This study presented a clear and structured framework for data preprocessing in machine learning, using a multilingual healthcare dataset as an example. Preprocessing is not just a small or routine task. It is very important for data quality, model performance, fairness, and clarity.

The techniques used in this study had three main goals: fixing structure, reducing noise, and making data ready for models. Structural changes like language translation and column splitting improved meaning, helped teams work across languages, and gave more detailed features. For example, changing Malay data to English made tokenization and meaning more consistent. Splitting fields allowed time or category details to appear that would stay hidden.

To reduce noise and keep data honest, missing values were marked and not deleted. This kept the chance to learn from missing parts, which is important in healthcare. Removing outliers, guided by stats and expert knowledge, saved models from errors from strange points. Finally, categorical data was turned into machine-readable and fair format using one-hot encoding. This avoided false order problems.

Together, these steps created a strong and repeatable preprocessing process that made the data more ready for models and useful in real cases. The results show that preprocessing is key for strong, high-quality, and clear machine learning models.

In the future, some things can make this process better. Techniques like Principal Component Analysis (PCA) or t-SNE can help with large and complex data. Scaling features, using normalization or standardization, help models that depend on distance. Also, advanced ways to fill missing values, using expert knowledge, can give better accuracy. Adding explainable AI tools like SHAP or LIME into preprocessing can make things clearer, especially in sensitive cases.

In short, preprocessing is not a small step before modeling but a strong and important part that shapes fairness, accuracy, and repeatability. The pipeline in this study is a good start that can work in many fields, especially where strong and clear machine learning solutions are needed.

## ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

## FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

## AUTHOR CONTRIBUTIONS

Md. Abu Nishan: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;  
Choo-Yee Ting: Project Administration, Supervision, Writing – Review & Editing;  
Chih-Yang Pee: Project Administration, Supervision, Writing – Review & Editing;  
Helmi Zakariah: Project Administration, Supervision, Writing – Review & Editing;  
Thanh-Trung Nguyen: Project Administration, Writing – Review & Editing.

## CONFLICT OF INTERESTS

No conflicts of interest were disclosed.

## ETHICS STATEMENTS

Our publication ethics follow the Committee of Publication Ethics (COPE) guidelines. <https://publicationethics.org/>

## DATA AVAILABILITY

Derived data supporting the findings of this study are available from the corresponding author on request.

## REFERENCES

- [1] P. Dhawas, A. Dhore, D. Bhagat, R. D. Pawar, A. Kukade, and K. Kalbande, "Big data preprocessing, techniques, integration, transformation, normalisation, cleaning, discretization, and binning," in *Advances in business information systems and analytics book series*, 2024, pp. 159–182. doi: 10.4018/979-8-3693-0413-6.ch006.
- [2] S. Dong, S. Sahri, and T. Palpanas, "Data Quality Awareness: A Journey from Traditional Data Management to Data Science Systems," *ACM SIGMOD Record*, vol. 55, no. 2, pp. 23–24, Jul. 2026, doi: 10.1145/3829088.3829090.
- [3] A. M. Sharifnia, D. E. Kpormegbey, D. K. Thapa, and M. Cleary, "A primer of data cleaning in quantitative research: handling missing values and outliers," *Journal of Advanced Nursing*, vol. 82, no. 1, pp. 970–975, Jan. 2026, doi: 10.1111/jan.16908.
- [4] C. E. Morr, M. Jammal, H. Ali-Hassan, and W. Ei-Hallak, *Machine learning for practical decision making*. 2022. doi: 10.1007/978-3-031-16990-8.

- [5] R. Thirunavukarasu, G. P. D. C, G. R, M. Gopikrishnan, and V. Palanisamy, "Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review," *Computers in Biology and Medicine*, vol. 149, p. 106020, Aug. 2022, doi: 10.1016/j.combiomed.2022.106020.
- [6] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Publisher Correction: Principal component analysis," *Nature Reviews Methods Primers*, vol. 3, no. 1, Mar. 2023, doi: 10.1038/s43586-023-00209-y.
- [7] H. Jeon, and S. Oh, "Hybrid-Recursive feature elimination for efficient feature selection," *Applied Sciences*, vol. 10, no. 9, p. 3211, May 2020, doi: 10.3390/app10093211.
- [8] M. A. K. Akhtar, M. Kumar, and A. Nayyar, "Transparency and Accountability in Explainable AI: Best Practices," in *Studies in systems, decision and control*, 2024, pp. 127–164. doi: 10.1007/978-3-031-66489-2\_5.
- [9] A. N. Prasad, "Data quality and preprocessing," in *Apress eBooks*, 2024, pp. 109–223. doi: 10.1007/979-8-8688-1023-7\_3.
- [10] K. M. Rao, G. Saikrishna, and K. Supriya, "Data preprocessing techniques: emergence and selection towards machine learning models - a practical review using HPA dataset," *Multimedia Tools and Applications*, vol. 82, no. 24, pp. 37177–37196, Mar. 2023, doi: 10.1007/s11042-023-15087-5.
- [11] Md. Hossain, A. Devnath, and P. Karmokar, "Handling missing values and outliers in advanced data pre-processing: An enhancement of diabetes classification accuracy," *Research Square (Research Square)*, Sep. 2023, doi: 10.21203/rs.3.rs-3364064/v1.
- [12] Y. Gong, G. Liu, Y. Xue, R. Li, and L. Meng, "A survey on dataset quality in machine learning," *Information and Software Technology*, vol. 162, p. 107268, Jun. 2023, doi: 10.1016/j.infsof.2023.107268.
- [13] M. K. Dahouda, and I. Joe, "A Deep-Learned embedding technique for categorical features encoding," *IEEE Access*, vol. 9, pp. 114381–114391, Jan. 2021, doi: 10.1109/access.2021.3104357.
- [14] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y.-I. Cho, "Effective methods of categorical data encoding for artificial intelligence algorithms," *Mathematics*, vol. 12, no. 16, p. 2553, Aug. 2024, doi: 10.3390/math12162553.
- [15] F. Dominguez, J. Unger, M. Traube, B. Mant, C. Ertler, and W. Lechner, "Encoding-independent optimization problem formulation for quantum computing," *Frontiers in Quantum Science and Technology*, vol. 2, Sep. 2023, doi: 10.3389/frqst.2023.1229471.
- [16] C. K. Williams I., "Naive bayes classifiers and one-hot encoding of categorical variables," *arXiv (Cornell University)*, Apr. 2024, doi: 10.48550/arxiv.2404.18190.
- [17] F. Clerici and N. Nobani, "Categorical variable encoding methods for tabular data: a benchmarking study," *International Journal of Data Science and Analytics*, vol. 22, no. 1, Feb. 2026, doi: 10.1007/s41060-025-00886-w.
- [18] Ch. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, p. 100164, Jan. 2023, doi: 10.1016/j.dajour.2023.100164.
- [19] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," in *Advances in intelligent systems and computing*, 2018, pp. 511–518. doi: 10.1007/978-981-10-7563-6\_53.
- [20] D. Wu, X. Ma, and D. L. Olson, "Financial distress prediction using integrated Z-score and multilayer perceptron neural networks," *Decision Support Systems*, vol. 159, p. 113814, May 2022, doi: 10.1016/j.dss.2022.113814.
- [21] B. L. Ortiz *et al.*, "Data preprocessing techniques for AI and Machine learning readiness: Scoping Review of wearable sensor data in cancer care," *JMIR Mhealth and Uhealth*, vol. 12, p. e59587, Aug. 2024, doi: 10.2196/59587.
- [22] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, pp. 238–248, Jan. 2022, doi: 10.1016/j.susoc.2022.03.001.

## BIOGRAPHIES OF AUTHORS

|                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | <p><b>Md. Abu Nishan</b> is a data analyst, researcher, and IT consultant with over 10 years of experience, including expertise in banking MIS and data solutions. He is an Oracle Certified 11g Database Administrator (OCA, OCP- DBA). He obtained his bachelor's degree from the University of Calcutta, India. He holds an MSc in Computer Science &amp; Engineering (MSCSE) from United International University (UIU), Bangladesh. He is pursuing a PhD in Information Technology (Data Science) at Multimedia University (MMU), Malaysia, focusing on healthcare risk prediction, especially Chronic Disease. He can be contacted at email: 1221404588@student.mmu.edu.my.</p> |
|    | <p><b>Choo-Yee Ting</b> is a Professor at the Faculty of Computing and Informatics and Deputy Director, Government Liaison &amp; Industry Engagement, Multimedia University (MMU), Cyberjaya, Malaysia. He works on predictive analytics and Big Data and got his PhD (Information Technology), Specialization: Artificial Intelligence from Multimedia University in 2009. A Microsoft Research Fellow (2002), he has led national projects and won two Big Data prizes. He is also a certified trainer, consultant, and data scientist for Malaysia's COVID-19 vaccine programme. He can be contacted at email: cyting@mmu.edu.my.</p>                                              |
|   | <p><b>Chih-Yang Pee</b> is an Assistant Professor at the Faculty of Computing and Informatics, Multimedia University (MMU), Cyberjaya, Malaysia. He works in statistical image analysis, pattern recognition, and moment invariants. He got his PhD in Mathematics from the University of Malaya in 2013. He has published journals like Pattern Recognition Letters and IEEE Access and leads a MOHE-funded project on discrete moment invariants. He also trains industry people in data science and Microsoft MCP. He can be contacted at email: cypee@mmu.edu.my.</p>                                                                                                             |
|  | <p><b>Helmi Zakariah</b> is the Chief Executive Officer of Hayat Technologies Sdn Bhd. With a base in medicine and global health, his skill covers Big Data Analytics, Machine Learning, and global health policy. He has worked with government groups, WHO, UN, NGOs, and poor groups, which has grown his skill in advocacy, partnerships, and talks. Outside work, he likes to travel alone and read books on behavioral economics. He can be contacted at email: drhelmi@myhayat.my.</p>                                                                                                                                                                                         |
|  | <p><b>Thanh-Trung Nguyen</b> is a lecturer and Dean at the Faculty of Information Technology - Communications, Dong Nai Technology University, Dong Nai, Vietnam. He received his M.S. degrees (2008) and Ph.D. (2018) in Computer Science at University of Information Technology, Vietnam National University – HoChiMinh City, Vietnam. His research interests include Data Mining, Big Data, Blockchain, Face Recognition, Camera Cross-Tracking, IoT and AI-Robotics. He can be contacted at email: nguyenthanhtrung@dnvu.edu.vn.</p>                                                                                                                                            |