Stress Detection using Naive Bayes and Decision Tree Classifiers

N Krishna Chowdary [1*] Pathuri, Shaik Rasool Basha [1*], Taraswi Jampana [1*], Vishnu Sathvik Reddy [1*], Meghanadh Reddy [1*], Dr M Kavitha [1]

[1*] Student, [1] Associate Professor
Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation (KLEF),
Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302
Email: mkavita@kluniversity.in

DOI : 10.48047/IJFANS/11/Sp.Iss5/062

Abstract. A large percentage of people deal with the ubiquitous problem of stress in our ever-busier lives, which includes both emotional and physical strains. Regrettably, a lack of awareness causes many people to fail to identify the symptoms of stress. Early detection of stress is essential because, if left unchecked, its effects can lead to serious health issues that progressively erode one's physical and mental health. Automated methods help to lower the risks related to stress that go untreated in addition to making stress detection easier. Although a number of models, including the multi-attention model, the Factor Graph model, and the Personal Knowledge model, have been developed for social media blogs, their efficacy has been limited because they have mainly examined single-line text. To overcome this constraint, a Decision Tree and Naive Bayes classifiers has been presented, with the goal of identifying stress from multi-line text data in the dataset that includes user-generated content that is both stressed and unstressed. The method has practical potential as evidenced by the promising accuracy of the experimental results.

Keywords: Stress Detection, Naïve Bayes, Decision Tree, Machine Learning

1. Introduction

The identification and classification of stress has become a significant difficulty in recent times. There are three types of stress: acute, episodic acute, and chronic stress. Stress is described as a condition of mental or emotional strain. It affects several body systems, including as the immune system, metabolism, and memory. Unfortunately, people who experience stress are frequently ignorant of their situation due to the lack of trustworthy stress-detecting technologies [1]. Effective treatment of stress requires early identification. The importance of stress detection in modern culture is shown by the reliance of stress detection techniques on textual, visual, and social indicators [2]. Early detection of stress has a major influence on treatment plans and patient outcomes.

In addition to being difficult and time-consuming, manual stress detection is also error prone. Therefore, it is imperative to have automated stress detection techniques. This work provides an overview, a problem description, objectives, a study of current systems, importance, and limitations of the Stress, which is basically a sensation of physical or emotional pressure, can have many causes, such as anxiety, rage, or irritation. Stress is a frequent problem, but if ignored, it can have serious consequences that could result in fatal illnesses including diabetes, cancer, heart disease, and depression. Direct questioning is a key component of traditional stress detection techniques, and if it is ignored, it can lead to depression. Since stress is not always obvious, it is critical to comprehend the causes that lead to it [3]. Sleep issues, light-headedness, worry, exhaustion, mood fluctuations, sickness, tense muscles, and shaking are common indicators of stress [4]. Stress may come from a variety of places, including the workplace, education, family, and social relationships. When under pressure, the degree of stress can quickly increase. Chronic mood disorders or addiction are examples of long-lasting stress that can have severe effects over a lengthy period of time [5]. Short-term stress, on the other hand, can typically be resolved quickly and arises from everyday obstacles like traffic or interpersonal issues. Because social media is so widely used as a platform for emotional expression, a lot of study has been done in this area, yet the findings are frequently disappointing [6]. Enhancing efficiency requires identifying and overcoming the restrictions in stress detection via study. Furthermore, it is imperative to investigate various stress indicators and their impact on individuals. In order to better understand this intricate phenomenon, a text segment is used as an input for stress detection in this study.

A machine learning approach uses a forecasting model that is built using historical data to predict the overall outcome when it receives user input [7]. How well the output is estimated depends on the amount of information used, since a large dataset makes it possible to create a model that quantifies the result. Machine learning has gained importance. Because machine learning can accomplish tasks that would be too complicated for a human to perform directly, it is thought to be necessary [8]. The machine must look for configurations and react to them correctly. An argument with a friend or a traffic jam are examples of everyday activities that can quickly relieve stress. These kinds of stress are referred to as acute or transient stress. Long-lasting stress, like that caused by alcoholism or mood disorders, is referred to as chronic stress [9]. The scientific community should identify and address the implications in stress detection for ongoing advancements. Furthermore, a variety of results ought to be acquired in order to ascertain the individual's level of impact. A few lines of text were used as input to identify stress to ascertain all of this.

2. System Design

In figure 1, the architecture of the Stress Detection Architecture describes the flow of the project where the dataset is loaded, and the data pre-processing step is carried out along with the feature selection. Then the dataset is spitted into test and train which carried out the machine learning techniques are carried out. Here we use two algorithms to calculate the accuracy of the system, namely Decision Tree and Naïve Bayes algorithms, where it finally results whether a person contains stress or not.

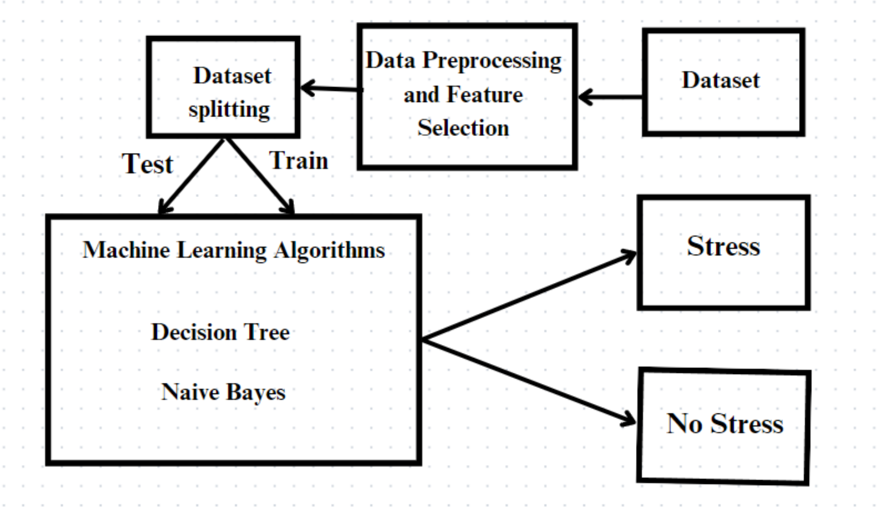


Figure 1 Stress Detection Architecture

The symptoms of stress can vary from person to person, and it is not imperative for everyone to experience the same manifestations [10]. However, if an individual's usual behaviour shifts towards abnormal patterns, it becomes crucial to observe and take necessary steps to identify the underlying issue. Stress can manifest in both physical and mental symptoms within the body, with common physical indicators encompassing aches and pains, sleep disturbances, high blood pressure, and digestive issues.

Challenges in stress detection

- The classification of the different words into different categories, because all words are not similar.
- Identifying the level of the stress means at which level the person is having stress.
- Word embedding have a significant impact on identifying the accuracy.

3. Methodology

Detecting stress levels from social media blogs can be difficult since stress can appear in several ways and individuals may not always directly state their stress levels in their articles. The proposed Decision tree algorithm is a machine learning technique that attempts to determine whether a person is stressed based on the content of their social media blogs.

Decision Tree Classifier

The decision tree algorithm is a form of supervised learning algorithm that makes predictions using a tree-like model. In this scenario, the algorithm would be trained on a dataset of social media blogs that had been labelled as stress-inducing or stress-free. The system would analyze numerous blog post features, such as the language used, the length of the post, and the presence of specific keywords, to determine whether a new post indicates stress or not.

Naive Bayes Classifier

Based on Bayes' theorem, the Naive Bayes Classifier is a probabilistic machine learning algorithm. It works especially well for text classification issues, like sentiment analysis and spam detection, but it can also be applied to other kinds of data. Based on the likelihood of the observed data (features), the Bayes theorem determines the probability of a hypothesis (class label).

Data Collection

Publicly available data set with 2838 multi-line stressed words in the training dataset and 715 multi-line stressed words in the testing dataset is collected from Kegel for implementation.

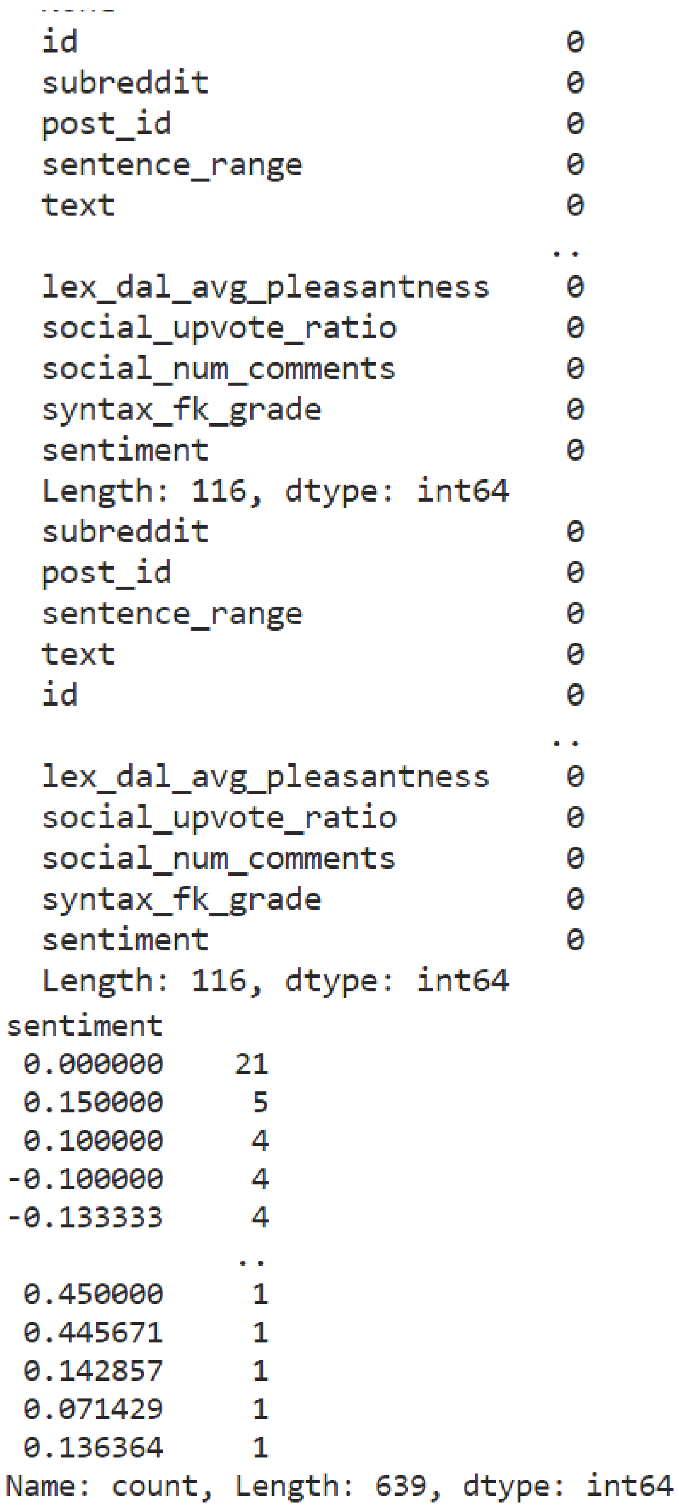
Data Preprocessing

The pre-processing is essential to handle the missing values and to address inconsistencies. Dataset is processed and removed all the null values and duplicate values.
chi-square (χ^2) test is a valuable tool for assessing the association between categorical features and a categorical target variable, which can be part of the process of feature selection in data analysis and machine learning.

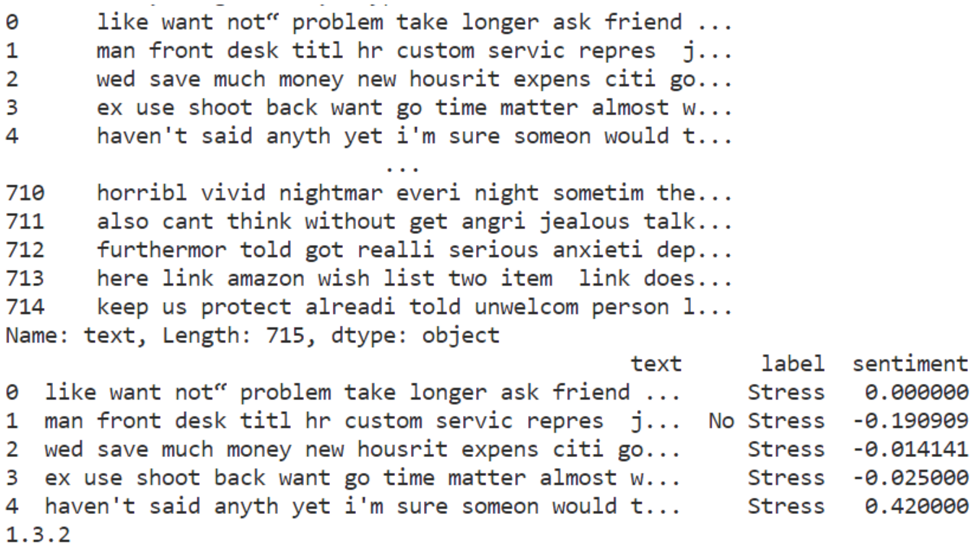
Dataset is divided into two parts training part and testing part in 80:20 ratios. Training part of the dataset is used to train the designed model and testing part of the dataset is used to evaluate the performance of the model.

3. Results Discussion

The project is implemented using Jupiter notebook. The dataset is loaded into environment and for further processing. Dataset is pre-processed and avoided null values and duplicate values. The features of the dataset are represented in figure 2.



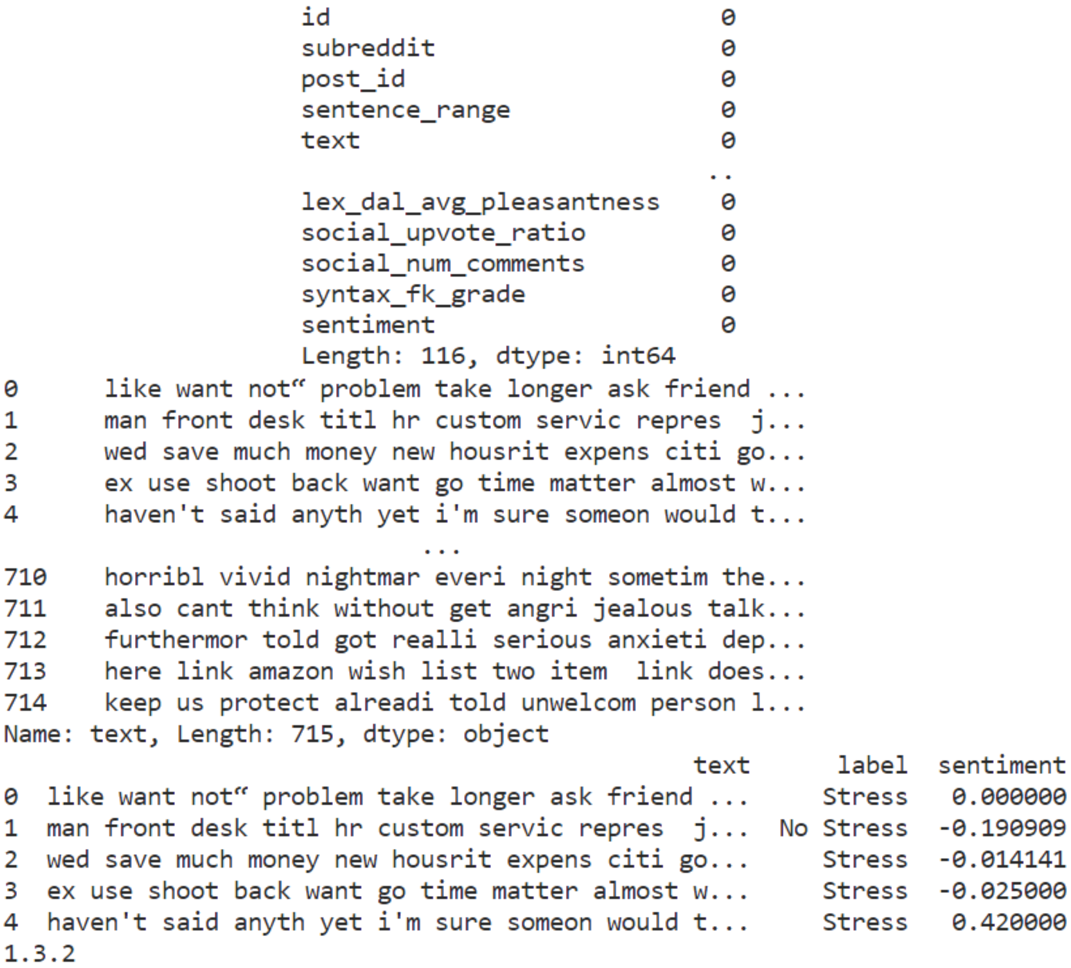
```
  id                         0
  subreddit                  0
  post_id                    0
  sentence_range             0
  text                       0
                            ..
  lex_dal_avg_pleasantness   0
  social_upvote_ratio        0
  social_num_comments        0
  syntax_fk_grade            0
  sentiment                  0
  Length: 116, dtype: int64
  subreddit                  0
  post_id                    0
  sentence_range             0
  text                       0
  id                         0
                            ..
  lex_dal_avg_pleasantness   0
  social_upvote_ratio        0
  social_num_comments        0
  syntax_fk_grade            0
  sentiment                  0
  Length: 116, dtype: int64
sentiment
 0.000000    21
 0.150000     5
 0.100000     4
-0.100000     4
-0.133333     4
             ..
 0.450000     1
 0.445671     1
 0.142857     1
 0.071429     1
 0.136364     1
Name: count, Length: 639, dtype: int64
```



```
0      like want not“ problem take longer ask friend ...
1      man front desk titl hr custom servic repres  j...
2      wed save much money new housrit expens citi go...
3      ex use shoot back want go time matter almost w...
4      haven't said anyth yet i'm sure someon would t...
                              ...
710    horribl vivid nightmar everi night sometim the...
711    also cant think without get angri jealous talk...
712    furthermor told got realli serious anxieti dep...
713    here link amazon wish list two item  link does...
714    keep us protect alreadi told unwelcom person l...
Name: text, Length: 715, dtype: object
                                                text      label  sentiment
0  like want not“ problem take longer ask friend ...     Stress   0.000000
1  man front desk titl hr custom servic repres  j...  No Stress  -0.190909
2  wed save much money new housrit expens citi go...     Stress  -0.014141
3  ex use shoot back want go time matter almost w...     Stress  -0.025000
4  haven't said anyth yet i'm sure someon would t...     Stress   0.420000
1.3.2
```

Figure 2: Features of the dataset

Chi-square (χ^2) method is applied and it identifies the association between categorical features and a target variable. The features after applying this preprocessing mechanism is shown in figure 2. Results evidence that feature selection enhances the models performance in stress detection.



```
                     id                          0
                     subreddit                   0
                     post_id                     0
                     sentence_range              0
                     text                        0
                                                ..
                     lex_dal_avg_pleasantness    0
                     social_upvote_ratio         0
                     social_num_comments         0
                     syntax_fk_grade             0
                     sentiment                   0
                     Length: 116, dtype: int64
0      like want not“ problem take longer ask friend ...
1      man front desk titl hr custom servic repres  j...
2      wed save much money new housrit expens citi go...
3      ex use shoot back want go time matter almost w...
4      haven't said anyth yet i'm sure someon would t...
                              ...
710    horribl vivid nightmar everi night sometim the...
711    also cant think without get angri jealous talk...
712    furthermor told got realli serious anxieti dep...
713    here link amazon wish list two item  link does...
714    keep us protect alreadi told unwelcom person l...
Name: text, Length: 715, dtype: object
                                                text      label  sentiment
0  like want not“ problem take longer ask friend ...     Stress   0.000000
1  man front desk titl hr custom servic repres  j...  No Stress  -0.190909
2  wed save much money new housrit expens citi go...     Stress  -0.014141
3  ex use shoot back want go time matter almost w...     Stress  -0.025000
4  haven't said anyth yet i'm sure someon would t...     Stress   0.420000
1.3.2
```

Figure 3 Features after pre-processing

Decision tree and Naïve Bayes models are applied on the dataset on original dataset and a dataset with feature extracted. The performance of both models are tested in terms of accuracy metric. Table 1 shows the performance analysis of applied approaches and figure 4 represents its graphical representation.

Table 1 Accuracy Analysis of classifiers

Model	On original Dataset	After Feature Selection
Decision Tree	0.53	0.75
Naïve Bayes	0.51	0.80

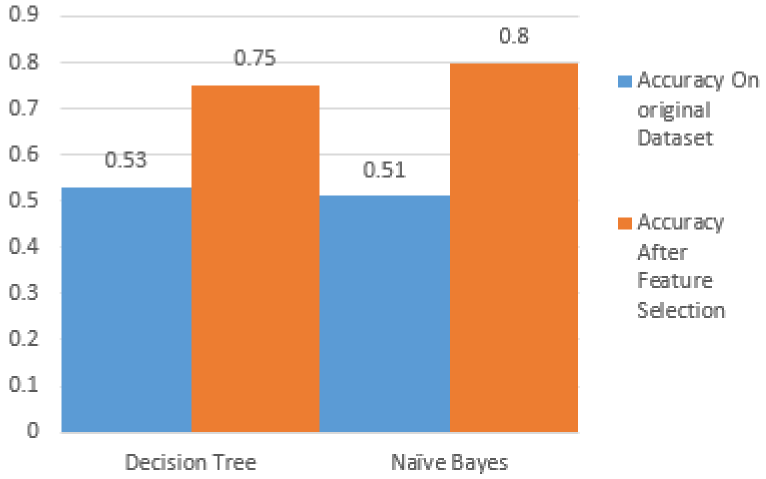


Figure 4 Graphical representaion of classifiers performance analysis

Conclusions

In this work two supervised machine learning approaches are applied on multi-line text data in the dataset, which is collected from kaggle platform. The dataset is pre-processed as per the requirement of the classifier. The dataset is divided into training and testing parts. The decision tree and naïve bayes classifiers are applied on the original dataset and the feature extracted dataset. The performance of the models are evaluated in terms of accuracy parameter. The results evidence that machine learning models are well suited for early stage detection of stress based on the words of people. By using the two methodologies, named Naïve Bayes and Decision Tree, we get the accuracy score with less accuracy. Later by applying Feature Selection and using Chi–Square test we increase the accuracy of our model, as the Chi–Square test is used for calculating Chi-square between each feature and the target and select the desired number of features with best Chi-square score.

Acknowledgement

Sincere thanks to the management of KL University for providing good infrastructure to do extensive research.

References

[1] Trotzek, M., Koitka, S., & Friedrich, C. M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering, 32(3), 588-601.

[2] William, D., & Suhartono, D. (2021). Text-based depression detection on social media posts: A systematic literature review. *Procedia Computer Science*, *179*, 582-589.

[3] Uban, A. S., Chulvi, B., & Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, *124*, 480-494.

[4] Elzeiny, S., & Qaraqe, M. (2018, October). Machine learning approaches to automatic stress detection: A review. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-6). IEEE.

[5] Deng, Y., Chu, C. H., Si, H., Zhang, Q., & Wu, Z. (2012). An investigation of decision analytic methodologies for stress identification. *International Journal on Smart Sensing and Intelligent Systems*, *6*(4), 1675-1699.

[6] Gedam, S., & Paul, S. (2021). A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, *9*, 84045-84066.

[7] Sharma, D., Kapoor, N., & Kang, S. S. (2020). Stress prediction of students using machine learning. *International* Journal of Mechanical and Production Engineering Research and Development, 10(3).

[8] Keshan, N., Parimi, P. V., & Bichindaritz, I. (2015, October). Machine learning for stress detection from ECG signals in automobile drivers. In 2015 IEEE International conference on big data (Big Data) (pp. 2661-2669). IEEE.

[9] Gupta, M., & Gupta, B. (2018, February). A comparative study of breast cancer diagnosis using supervised machine learning techniques. In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC) (pp. 997-1002). IEEE.

[10] Wadkar, K., Pathak, P., & Wagh, N. (2019). BREAST CANCER DETECTION USING ANN NETWORK AND PERFORMANCE ANALYSIS WITH SVM. Journal of Computer Engineering and Technology, 10(3), 75-86.

[11] Natarajan, K., B. Prasath, and P. Kokila. "Smart health care system using internet of things." Journal of Network Communications and Emerging Technologies (JNCET) 6.3 (2016).

[12] Lin, H., Shao, J., Zhang, C., & Fang, Y. (2013). CAM: cloud-assisted privacy preserving mobile health monitoring. IEEE Transactions on Information Forensics and Security, 8(6), 985-997.