Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220626

Prediction of Skin Cancer Using Pre-Trained Language Models from Patient Symptoms

D.Deepa[1], R.Yaswanth[1] and C.Suganth[1]

[1]Department of Computer Science and Engineering, Kongu Engineering College, Erode, Tamilnadu, INDIA

deepa@kongu.ac.in

Abstract. Automatic feature extraction and processing of greater data is now possible because of advances in Deep Learning. To pre-train from a wider corpus and comprehend the language feature for sentiment classification work, transformers Generalized Autoregressive Pre-training for Language Understanding and Bidirectional Encoder Representations from Transformers (BERT) have been proposed. These language models learn the context in both ways. In the proposed work, we have examined and tested our text dataset of skin cancer cases using the BERTbase model. When determining whether a patient's symptoms are compatible with cancer or not the model has a 97.3 percent accuracy rate.

Keywords: sentiment classification, bidirectional, encoders, transformers, reviews.

1. Introduction

Sentiment analysis is an NLP (Natural language processing) that determines an opinion represented a sentence, that can be found in a group of online reviews, tweets, blogs, and other forums. A polarity classification that can range from biclass to fine granularity is the result of the sentiment analysis task. Some examples are extremely negative, negative, neutral, positive, and extremely positive.

Sentiment analysis is an activity that helps firms develop by tracking and understanding client input on product sentiment. Customers may communicate their opinions and feelings more freely than previously thanks to the emergence of social media. Manually categorizing millions of daily feedbacks from stock traders and online retail users for sentiment analysis is not a simple task. This type of analysis deals with urgent real-time concerns and should assist individuals in taking immediate action. In addition, participants in the process spent sixty-five percent of their time identifying the sentiment based on the volume of the remark. To fulfill the work, a cost-effective autonomous sentiment analysis model that is context-aware is required.

Later, deep learning algorithms revolutionized data preparation by reducing defects and inadequacy. Deep learning methods automate word embedding, which is generally pre-trained on a text corpus based on co-occurrence data, reducing the need for manual preparation. Word2Vec and GloVe like models provide pre-trained word representation. Deep Learning models with diverse architectures, such as CNN, RNN, LSTM, and Bi-LSTM, learn context better. These context-free pre-trained word embeddings are used as language rules. It is feasible to use the acquired probability distributions to comprehend more languages.

DOI: 10.54254/2755-2721/2/20220626

2. Bert pre-trained model

Language models have the drawback of only learning from left- or right-hand settings, yet language is a two-way street. BERT, Google's language representation model, surpasses the competition in 11 natural language processing tasks because that was trained on larger data [1]. BERT has several faults, including random 15% masking and predictions. BERT was pre-trained in a vast corpus of unlabeled text containing 3,300 million words, which includes the entirety of Wikipedia and Book Corus. The BERT model is a two-way model. The Random Masking Pre-Training Model is used by BERT. Bidirectionality is the ability of BERT to read both the left and right sides of a word simultaneously. BERT is pre-trained on two separate NLP tasks using these bidirectional capabilities. The first job is masking, which involves hiding a word in a sentence and then asking the computer to infer the word that has been hidden (masked) from the context of the hidden word. The second test, called Next Sentence Prediction, asks the user to determine if two supplied sentences are connected logically and sequentially or if their relationship is just random.

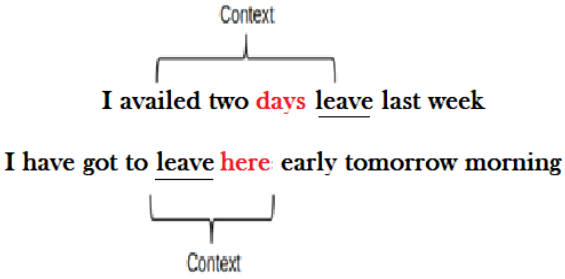

Figure 1. BERT learning the context.

When the meaning of the word "leave" in Fig.1. is exclusively dependent on the context to the left or right, at least one of the above two sentence will be erroneous. Before generating a forecast, one way is to consider both the left and right contexts. BERT accomplishes this. The transformer is the foundation for the BERT architecture. BERT is available in two flavors: BERT Large had 24 layers from a transformer block, 16 attention heads, and 350 million parameters, whereas BERT Base has 12 layers, 12 heads, and 120 million parameters. Encoder-only blocks make up all of these Transformer levels.

BERT has already completed two NLP tasks: masked language modelling and next sentence predictions. MLMs (Language Models with Masks) learn how words are connected. Only 15% of the words were randomly masked to avoid the model from focusing too much on one token or a location. The masked words are tagged as [MASK] token, which never occurs during fine-tuning.

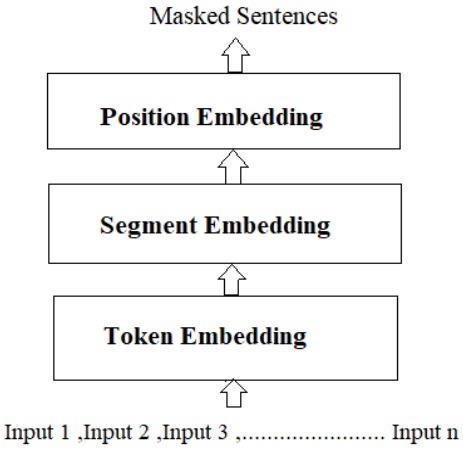

Figure 2. BERT masking.

DOI: 10.54254/2755-2721/2/20220626

In Fig.2 the input sentence will be given and Token Embedding takes place to pre-trained embeddings for various words known as token embeddings. To tokenize the text for these pre-train token embeddings, a method called WordPiece tokenization is utilized. And then Segment embeddings are essentially a vector representation of the sentence number. BERT is a model with absolute position embeddings, padding the inputs on the right rather than the left is typically recommended.

BERT has two versions: BERT Base, which has layers of transformer blocks, 12 attention heads, and 110 million parameters, and BERT Plus, which has 12 layers of transformer blocks, 12 attention BERT Large has 24 layers of transformer blocks, 16 attention heads, and 340 million variables. Encoder-only blocks make up each of these Transformer levels.

3. BERT fine-tunning

The intuition behind BERT is that the early layers learn generic linguistic patterns that have little relevance to the downstream task, while the later layers learn task-specific patterns. This intuition is in line with deep computer vision models, where the early layers learn in the case of facial recognition, general traits such as edges and corners are learned first, followed by specialized features such as eyes and noses. This intuition has been experimentally confirmed by another Google team, [2] One of their techniques is called partial freezing: they keep the early BERT layers frozen, i.e. fixed, during the fine-tuning process, and measure how much the performance on the downstream task changes when varying the number of frozen layers. They show that the performance on both MNLI and SQuAD tasks does not notably drop even when freezing the first 8 of the 12 BERT layers (i.e. tuning only the last 4). This finding corroborates the intuition that the last layers are the most task-specific, and therefore change the most during the fine-tuning process, while the early layers remain relatively stable. The results also imply that practitioners can potentially save compute resources by freezing the early layers instead of training the entire network during fine-tuning.

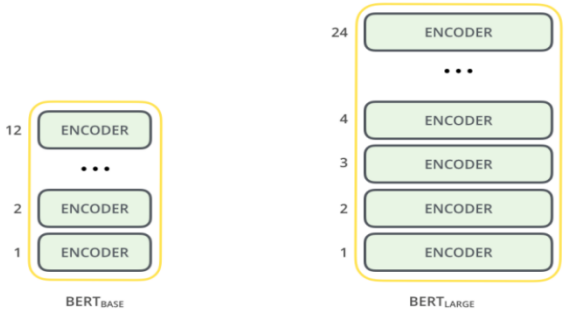

Figure 3. BERT types.

Google's BERT was a paradigm shift in natural language modeling, in particular, because of the introduction of the pre-training / fine-tuning paradigm: after pre-training in an unsupervised way on a massive amount of text data, the model can be rapidly fine-tuned on a specific downstream task with relatively few labels, because the general linguistic patterns have already been learned during pre-training. Fine-tuning BERT is "straightforward", simply by adding one additional layer after the final BERT layer and training the entire network for just a few epochs. The authors demonstrate strong performance on the standard NLP benchmark problems GLUE, SQuAD, and SWAG, which probe for different aspects of natural language inference, after fine-tuning for just 2-3 epochs with the ADAM optimizer, with learning rates between 1e-5 to 5e-5, a recipe that has been commonly adopted within the research community.

DOI: 10.54254/2755-2721/2/20220626

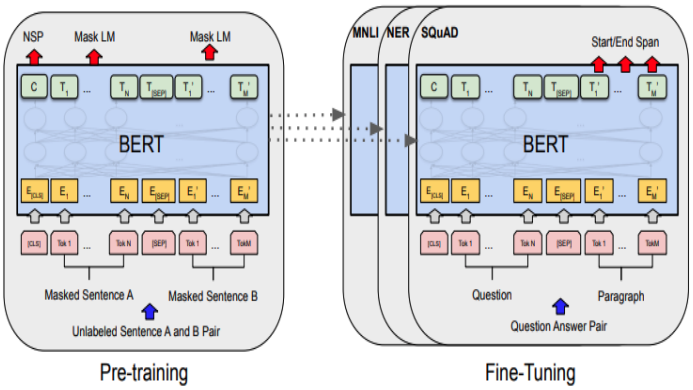

Figure 4. BERT and fine tuning [2].

4. Literature review

On a text classification task, researchers tested several BERT fine-tuning procedures [3]. They hypothesized and proved that language Model pre-training helps to acquire more contextual language information and achieve cutting-edge performance on several downstream NLP tasks. Domain-specific language representation model pre-trained on large-scale biomedical corpora. It is based on the domain of the medical field [4]. Domain-specific BERT depends on a particular domain. Deep learning has aided the creation of excellent biomedical text mining algorithms, which have gained appeal among researchers.

Researchers used transfer learning with BERT model for fine-grained sentiment classification task using SST dataset and showed outperforming results[5] [6] predict the Aspect-based sentiment analysis for the given aspects or targets by introducing a context-aware embeddings layer that contains the most correlated information for the selected context. By monitoring customers' views and actions, social media has an impact on their choices. Monitoring consumer loyalty and attitude towards businesses or goods via social media is a useful technique to gauge client loyalty [7]. The natural next marketing sector is social media. In terms of digital marketing, Facebook currently reigns supreme, closely followed by Twitter [8]. One year into the COVID-19 epidemic, and one of the world's longest lockdowns has occurred, Natural language processing techniques were used to decipher general sentiment, which will aid the government in understanding its reaction. The results yielded 85.99% accuracy has occurred.

Proposed an approach for the transmission of sentiment features from BERT output with context-aware embeddings is managed by the gating method. On the SentiHood and SemEval-2014 datasets, the suggested model obtained cutting-edge test F1 results of 88.0 and 92.9, respectively[9]. Stock information in relationship with opinion examination of information features and Twitter tweets information, to anticipate the future pace of a stock of interest. The Model uses three machine algorithms to predict: auto-regressive, Long short-term memory, and linear regression. To forecast the future pace of a stock of interest, the intention is to use authentic stock data in conjunction with opinion analysis of information characteristics and Twitter tweets data [10]. The customer comments were collected from Amazon and Flipkart data to review the product. With extensive process descriptions, a generic approach for categorizing sentiment polarity is presented. Online product reviews from Amazon.com were used in this investigation. Experiments on classification at the sentence and review of the level have predicted finding results. Finally, we discuss our plans for sentiment analysis in the future.

Created an ensemble of BERT models and distilled the information using a dataset of self labeled movie reviews [11]. The models were developed and achieved superior results thanks to BERT's pooling layer architecture.[12]. Outlined a strategy for enhancing the effectiveness of current text classification algorithms in fields with robust linguistic semantics. A generic and a domain-specific word embedding are combined into a domain-adapted embedding by the domain adaptation layer, which

DOI: 10.54254/2755-2721/2/20220626

learns weights. The generic encoder + classifier framework is then utilized to execute a downstream job, such as classification, using the DA word embeddings as inputs [13]. Researchers Presented an innovative strategy to enhance the performance of sentiment classification by raising the standard of sentiment lexicons [14]. This is achieved by keeping note of the sentences that were incorrectly predicted and using those as supervision. The polarity score of one word is updated while the search for fresh sentiment words is balanced by an exploration-exploitation method [15].

5. Proposed work

The World Wide Web is a large database of organized and unstructured information. Opinion gathering or Sentiment analysis involves a classification task which needs a large volume of training data. But the reviews we obtain may be unstructured [16] [17]. Fig 4 shows the architecture diagram for the BERT fine-tuning and transfer learning using BERT respectively. First Load the datasets which are pre-labeled in the excel format which has two columns of review and sentiment. Then the learning process starts with preprocessing the dataset to improve the accuracy of the system. WordPiece tokenizes each word into subword units called word pieces after pre-tokenizing the text into words by splitting punctuation and with whitespaces [18].

Location embeddings are the vector representations of the position of a word inside a phrase. Token embeddings, although being a very essential and valuable embedding, do not provide information on the token's position in a phrase. So another embedding called position embeddings is utilized to solve it.

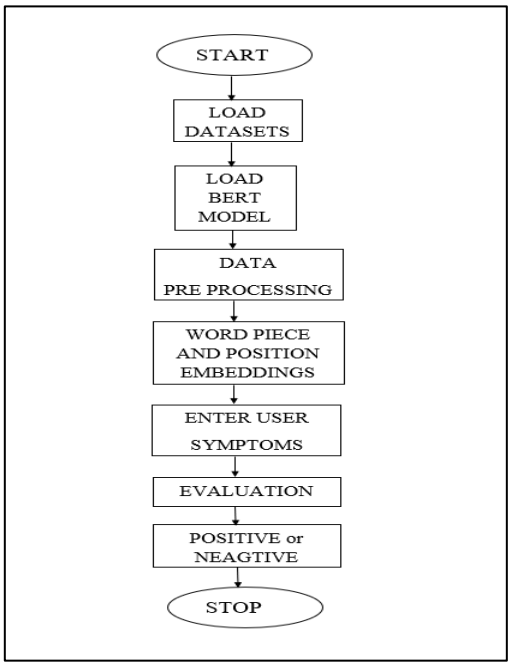

Figure 5. Pre-training using BERT.

6. Transformers

BERT's underlying model is based on the Transformer architecture. The Transformer architecture is a paradigm that relies on attention over the sequence rather than recurrent connections. Multiple attention chunks make up the Transformer. Each block applies attention to the sequence and alters the input using linear layers. The transformers are the first transduction model to calculate representations of its input and output using self-attention rather than sequence-aligned RNN. The conversion of input sequences into output sequences is known as transduction. The transformer's goal is to completely manage the dependencies between input and output using attention and repetition [19][20].

DOI: 10.54254/2755-2721/2/20220626

7. Dataset

The collection includes approximately 2,100 pieces of data and numerous reviews called from Twitter comments and pre-labeled datasets. For training and testing, 1050 review samples are used. Because the reviews are pre-labeled with the sentiment, the technique for evaluating sentiment was used. The code was written entirely in Colab, a strong Python programming environment.

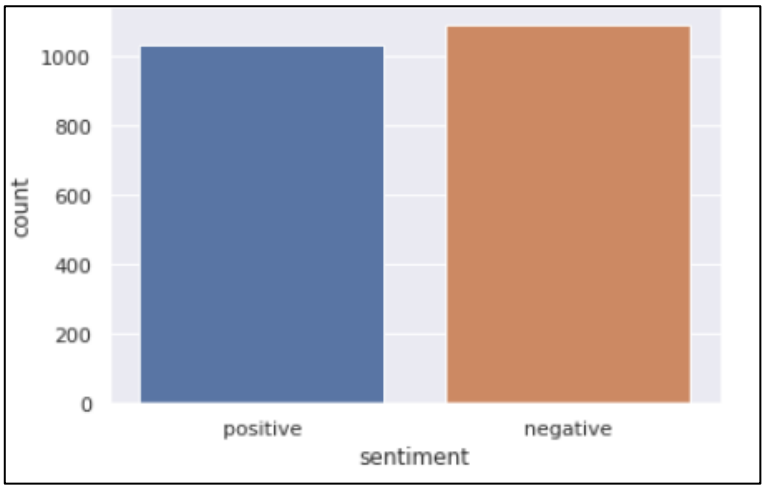

Figure 6. Number of positive and negative reviews.

8. Result and discussion

Table 1 reveals that BERT has the best accuracy for the data set. XLNET is a deep learning model for word embedding that has been pre-trained, but only learns the text from the left direction to right direction. The other two pre-training language models are bidirectional and learn contexts from the largest corpus. BERT also employs transformers and attention heads to better capture the settings. By masking words and learning the text, XLNET covers BERT discrepancy. As a result, pre-trained language models classify reviews better than pre-trained embedding mode.

Table 1. Accuracy – deep learning and pre-trained models.

MODEL	ACCURACY
RNN	71.1
XLNET	95.6
BERT	97.3

9. Loss and accuracy

Each iteration of the training procedure calculates the loss and accuracy. Each time, the accuracy improves while the loss reduces. We run an average epoch value of 2 epochs. The loss was 45.71 percent and the accuracy was 80.06 percent in iteration 1, but the loss was decreased to percent 15.53 and the accuracy increased to 97.34 percent, as shown in figure 1.8.

```
Epoch 1/2
54/54 [==============================] - 128s 2s/step - loss: 0.4571 - accuracy: 0.8006 - val_loss: 0.1740 - val_accuracy: 0.9499
Epoch 2/2
54/54 [==============================] - 95s 2s/step - loss: 0.1553 - accuracy: 0.9599 - val_loss: 0.0896 - val_accuracy: 0.9734
0.9733959436416626
```

Figure 7. Accuracy and loss calculation.

DOI: 10.54254/2755-2721/2/20220626

In Fig.7, the graph represents the accuracy and the loss comparison for each epoch, here the accuracy increases for each epoch and loss decreases, and then the training time also reduces from each epoch to epoch.

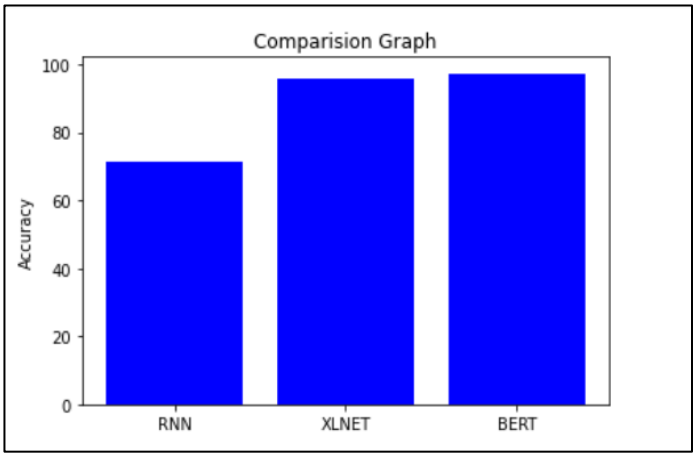

Figure 8. A RNN, BERT and XLNET accuracy.

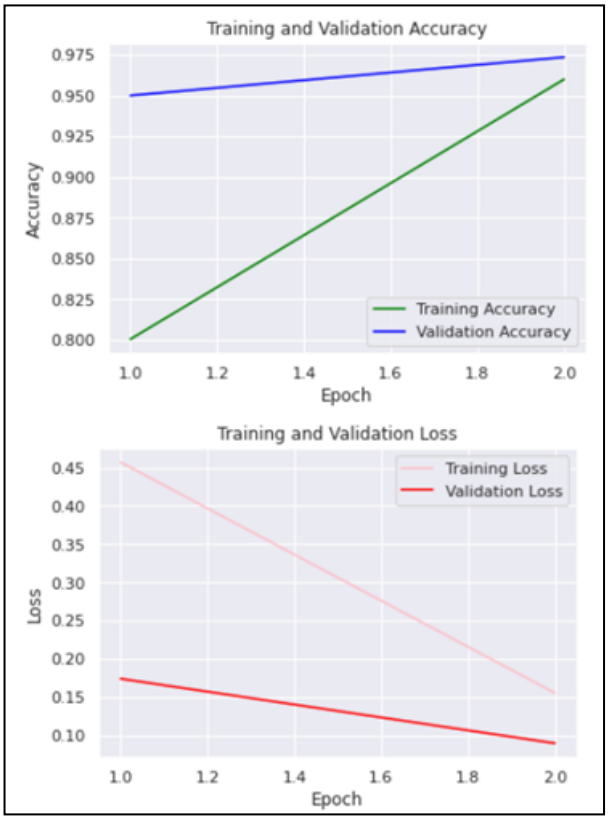

Figure 9. Accuracy and loss calculation.

10. Conclusion and future work

Pre-trained language models are designed to learn properties that are relevant in tasks such as question answering, next sentence predictions, sentiments analysis, named entities recognition, and so on. In our investigation, to comprehend the language's context, use the dependencies between the mask word and the different words. For a Skin cancer dataset, one of the models we tested outperformed BERT.

DOI: 10.54254/2755-2721/2/20220626

In future work, the model will be included a solution for the patient's symptoms when the result is positive.

References

[1] Devlin, J., et al., BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[2] Pavithra, E., Janakiramaiah, B., Narasimha Prasad, L. V., Deepa, D., Jayapandian, N., & Sathishkumar, V. E.. Visiting Indian Hospitals Before, During and After Covid. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems., 30(1), pp. 111-123,2020.

[3] Sun, C., et al. How to fine-tune BERT for text classification? in China national conference on Chinese computational linguistics. 2019. Springer.

[4] Lee, J., et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020. 36(4): p. 1234-1240.

[5] Munikar, M., S. Shakya, and A. Shrestha. Fine-grained sentiment classification using BERT. in 2019 Artificial Intelligence for Transforming Business and Society (AITB). 2019. IEEE.

[6] Li, X., et al., Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. IEEE Access, 2020. 8: p. 46868-46876.

[7] Jain, P.K., et al., Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. Journal of Ambient Intelligence and Humanized Computing, 2022: p. 1-13.

[8] Nezhad, Z.B. and M.A. Deihimi, Twitter sentiment analysis from Iran about COVID 19 vaccine. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 2022. 16(1): p. 102367.

[9] Singh, M., A.K. Jakhar, and S. Pandey, Sentiment analysis on the impact of coronavirus in social life using the BERT model. Social Network Analysis and Mining, 2021. 11(1): p. 1-11.

[10] Shah, B.K., et al. Sentiments Detection for Amazon Product Review. in 2021 International Conference on Computer Communication and Informatics (ICCCI). 2021. IEEE.

[11] Lehečka, J., et al. BERT-based sentiment analysis using distillation. in International Conference on Statistical Language and Speech Processing. 2020. Springer.

[12] Sarma, P.K., Y. Liang, and W.A. Sethares, Shallow domain adaptive embeddings for sentiment analysis. arXiv preprint arXiv:1908.06082, 2019.

[13] Xing, F.Z., F. Pallucchini, and E. Cambria, Cognitive-inspired domain adaptation of sentiment lexicons. Information Processing & Management, 2019. 56(3): p. 554-564.

[14] Sathishkumar V E, Changsun Shin, Youngyun Cho, "Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city", Building Research & Information, Vol. 49. no. 1, pp. 127-143, 2021.

[15] Sathishkumar V E, Youngyun Cho, "A rule-based model for Seoul Bike sharing demand prediction using Weather data", European Journal of Remote Sensing, Vol. 52, no. 1, pp. 166-183, 2020.

[16] Sathishkumar V E, Jangwoo Park, Youngyun Cho, "Using data mining techniques for bike sharing demand prediction in Metropolitan city", Computer Communications, Vol. 153, pp. 353-366, 2020.

[17] Sathishkumar V E, Yongyun Cho, "Season wise bike sharing demand analysis using random forest algorithm", Computational Intelligence, pp. 1-26, 2020.

[18] Sathishkumar, V. E., Hatamleh, W. A., Alnuaim, A. A., Abdelhady, M., Venkatesh, B., & Santhoshkumar, S. (2021). Secure Dynamic Group Data Sharing in Semi-trusted Third Party Cloud Environment. Arabian Journal for Science and Engineering, 1-9.

[19] Sathishkumar V E., Jangwoo Park, Youngyun Cho, "Seoul Bike Trip duration prediction using data mining techniques", IET Intelligent Transport Systems, Vol. 14, no. 11, pp. 1465-1474, 2020.

[20] Sathishkumar Easwaramoorthy., Sophia, F., & Prathik, A. (2016, February). Biometric Authentication using finger nails. In 2016 international conference on emerging trends in engineering, technology and science (ICETETS) (pp. 1-6). IEEE.