

Amharic event text classification from social media using hybrid deep learning

Amogne Andualem Ayalew¹, Melaku Lake Tegegne², Bommy Manivannan³, Tamilarasi Suresh⁴, Napa Komal Kumar⁵, Battula Krishna Prasad⁶, Tsehay Admassu Assegie⁷, Ayodeji Olalekan Salau^{8,9}

¹Department of Information Technology, School of Computing and Informatics, Mizan-Tepi University, Tepi, Ethiopia

²Department of Information Technology, College of Engineering and Technology, Injibara University, Injibara, Ethiopia

³Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle, India

⁴Department of Information Technology, St. Peter's Institute of Higher Education and Research, Chennai, India

⁵Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, India

⁶Department of Computer Science and Engineering, Koneru Lakhmaiah Education Foundation, Guntur, India

⁷School of Electronics Engineering, Kyungpook National University, Daegu, Republic of Korea

⁸Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria

⁹Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

Article Info

Article history:

Received Feb 7, 2024

Revised Oct 4, 2024

Accepted Nov 20, 2024

Keywords:

Convolutional neural network

Deep learning

Event classification

Long short-term memory

Word embedding technique

ABSTRACT

This study aims to develop a hybrid deep-learning model for detecting and classifying Amharic text. Various natural language applications, such as information extraction, event extraction, conversation, text summarization, and require an automatic event classification. However, existing studies focused on classification, giving little attention to the preprocessing and feature extraction techniques. To address this problem, this work proposed a hybridized deep learning-based Amharic social media text event classification model. The model consists of word-to-vector (Word2vec) word embedding techniques to capture the semantic and syntactic representation. Convolutional neural network (CNN) is used to extract short-length text features. Additionally, bidirectional long-short memory (Bi-LSTM) is used to extract features from long Amharic sentences and classify those events based on their classes. The dataset used for training and testing consists of 6,740 labeled Amharic text sentences, collected from social media. The result shows an accuracy of 94.8% in detecting and classifying Amharic text events.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tsehay Admassu Assegie

School of Electronics Engineering, Kyungpook National University

41566, Daegu, Republic of Korea

Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

With the increasing growth of the internet, automatic event classification is becoming highly important to social media users [1]. Event classification has emerged as the primary focus of natural language processing with artificial intelligence techniques. Natural language processing (NLP) is the sub-field of artificial intelligence concerned with the automatic manipulation of natural language which encompasses speech recognition, and text classification [2], [3]. Text events on social media describe personal issues, different disasters, sensitive issues, incidents, earthquake events, terrorist attacks, and so on are reported every day. Therefore, automatically and quickly detecting and classifying Amharic text events based on their type is a highly practical research area [4], [5].

The classification of text event type plays a prominent role in developing different NLP applications such as event extraction, information extraction, text classification, risk analysis, and disaster prediction [6]. Many scholars have proposed various event classification models for different natural languages such as English, Spanish, and Chinese. However, because of the morphological diversity of the Amharic language, we cannot use it directly for the Amharic language event classification model. Previously, numerous researchers used unsupervised machine-learning methods to construct a variety of event detection and categorization systems [7], [8]. To that end, various classic machine learning event classification models such as the k-nearest neighbor (K-NN) classification algorithm, support vector machine (SVM), naïve Bayes (NB) classification algorithm, decision tree, and others have been applied for text document classification [9]. However, those machine learning approaches, on the other hand, have a variety of drawbacks for text classification applications.

Recently, several works [10], [11] have introduced the use of various machine-learning techniques for natural language processing. Among these, some studies proposed Amharic text event classification. A study [12] proposed different research related to Amharic events by the applied method of rule-based and traditional machine learning feature selection methods to identify, classify, and extract events from unstructured texts.

Some researchers have conducted different studies in different languages on different event types such as financial events [13], drug abuse events [14], [15] biomedical events. Hence, event classification is a required task for different natural language processing tasks. Some works are proposed on Amharic events [16], [17] the authors use different traditional machine learning and rule-based to classify Amharic document event and non-event class types. Another work [18]–[21] also proposed extracting events from Amharic New articles and ontology-based event identification, with the event words collected manually from a different source and they used a shallow machine learning classifier called a maximum entropy to determine whether the sentence contains events or not. The work proposed an event extraction model from Amharic texts using deep learning approaches.

Recently, text event classification research has achieved more efficacy in European languages with traditional machine learning and deep learning. Most of the European language processing researchers performed bidirectional encoder representations from transformers to detect financial event classification. Convolutional neural network (CNN) models use convolutional layers and maximum pooling or max-over-time pooling layers to extract only local features of text. Long short-term memory is a recurrent neural network LSTM is an improved recurring neural network (RNN) architecture that uses a gating mechanism consisting of an input gate, forget gate, and output gate [22]–[24]. Also, BiLSTM captures semantic information of a text in a document by its preceding and following information in the text, while CNN is used to capture structure information from the local features [25]–[27].

In this study, to address the weakness (CNN and RNN), we proposed a CNN+Bi-LSTM hybrid model that classifies Amharic text events using a social media dataset. The study proposed a model that classifies Amharic text events. The model is designed to classify five different types of disaster events (such as conflict, traffic accident, fire, flood, and neutral events) from Amharic documents. Moreover, the study presented a comparative analysis of various deep learning approaches with word embedding techniques. The paper's organization is as follows: The section highlights the method. Section 3 discusses the result and section 4 concludes the study.

2. METHOD

The first preprocessing component has some sub-components such as tokenization, punctuation mark removal, character normalization, and word steaming. The second component is the Word embedding module which includes using one-hot encoding and developing alternative Word2vec algorithms such as continuous bag of word (CBOW) and Skip-gram, which is proposed by collecting Amharic text documents and selecting their parameters. In the third module event classification module, we designed a hybrid deep learning-based event classifier that classifies texts based on their pre-defined event type. Figure 1 shows the flowchart for the proposed study.

In this paper, different long and short Amharic documents are used for training and testing to design an event classification deep learning model. Our proposed Amharic text event classification model includes a total of 5 event categories and a total of 6,740 event texts. Documents are collected from different social media (such as Facebook, Twitter, and Telegram) datasets, which define 5 different types of Amharic events (conflict, traffic, flood, fire, and neutral). The dataset has been split into training, validation, and testing (80:10:10) splitting ratio. Based on the splitting ratio, 80% of the dataset was used for training, 10% of the dataset was used to validate the model, and 10% of the dataset was used to test the model. Figure 2 illustrates the distribution of the dataset. The model is evaluated using precision (P), recall (R), and F1-score (F1). Additionally, we compared other machine learning event classifier models (support vector machine (SVM),

conditional random fields (CRF), K-nearest neighbor (KNN), and naïve Bayes (NB)) with the term frequency-inverse document frequency (TFIDF) and bag of words (BoW) feature extraction techniques. Figure 2 shows the distribution of the dataset.

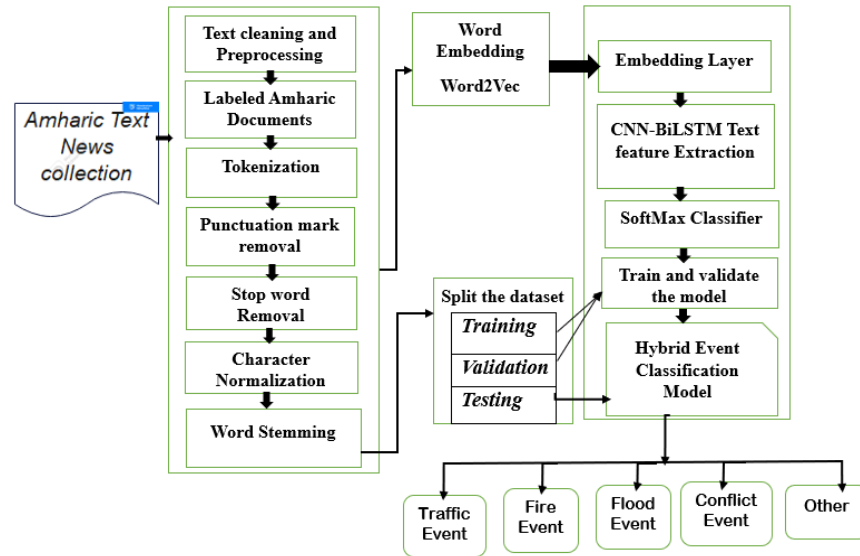


Figure 1. The proposed hybrid deep-learning text event classification

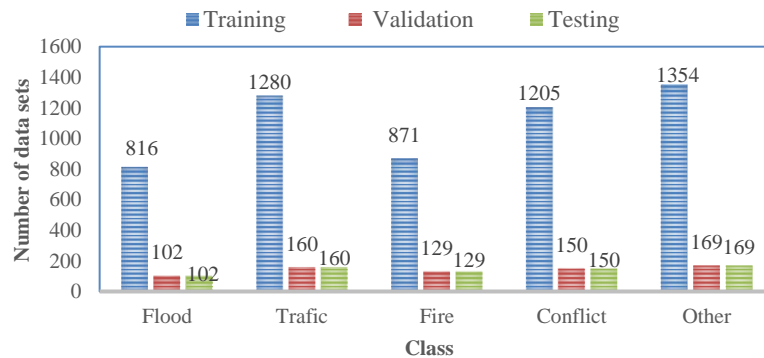


Figure 2. The distribution of the dataset

When compared to other event classification models, the random forest classifier with bag of words offers better-predicted results, as shown in Table 1 machine learning has several weaknesses, including feature engineering. To address this issue, we developed a new sophisticated deep learning event classification model called hybrid deep learning techniques. In CNN, the event classification model experiment uses different parameters including (embedding layer, text sequence length, convolutional kernel size, activation function, number of dense layers, optimization, and loss function).

Table 1. Parameters for CNN event classification model

Parameter	Result
The total length of the sentence	150
Kernel size	2, 3, and 4
Embedding size	300
epoch	20
Learning Rate	0.0001
CNN dropout probability	0.49
Optimizer	Adam

3. RESULTS AND DISCUSSION

In this paper, different long and short Amharic documents are used for training and testing to design an event classification deep learning model. Our proposed Amharic text event classification model includes a total of 5 event categories and a total of 6,740 event texts. Documents are collected on the widely used different social media (Facebook, Twitter, and Telegram) datasets, which define 5 different types of Amharic events (conflict, traffic, flood, fire, and neutral). The dataset has been split into training, validation, and testing (80:10:10) splitting ratio. Based on the splitting ratio, 80% of the total datasets were used for training, and 10% of the dataset was used to validate the model, the rest. Table 2 indicates the performance of various machine learning models.

However, we could not get better results using the CNN model because the dataset comprises long sequences by nature, resulting in poor model performance. Other RNN models were presented to increase the performance of an event classification model that uses different gates to capture context data from large text sequences. We implemented single LSTM and BiLSTM from this RNN model to collect contextual information texts in both forward and backward directions. As shown in Table 3, the hybrid of the two deep learning algorithms (CNN and BiLSTM) performed better results than other traditional machine learning and single deep learning models. This is because, CNN could capture local features of a text, and BiLSTM has the advantage of a global feature of a text, which can capture the features of a text including the context and semantics of a word.

Table 2. Performance of SVM, RF, and NB for Amharic text event classification

ML method	Feature extraction method	Accuracy
SVM	Bag of words	84.64%
	TFIDF	77.22%
RF	Bag of words	86.19%
	TFIDF	80.86%
NB	Bag of words	77.81%
	TFIDF	75.07%

Table 3. Deep learning event classification model results

Embedding layer	Model	Accuracy
Word2vec	CNN	89.4%
Word2vec	LSTM	87.6%
Word2vec	BiLSTM	92.1%
Word2vec	CNN - BiLSTM	94.8%

Figure 3 indicates that after the model-building process, the model is evaluated. Figure 3 shows that the training and validation start from 0.2 and finally reach 0.9 with epoch size 20. Whereas Figure 3 indicates the training and validation loss starts at 0.6 and goes down to 0.1 with the epoch size 20. This shows that our model learns more features from time to time and predicts effectively without model overfitting.

Figure 4 indicates that after the model-building process, the model is evaluated. Figure 4 also indicates the training and validation loss starts at 0.6 and goes down to 0.1 with the epoch size 20. This shows that our model learns more features from time to time and predicts effectively without model overfitting.

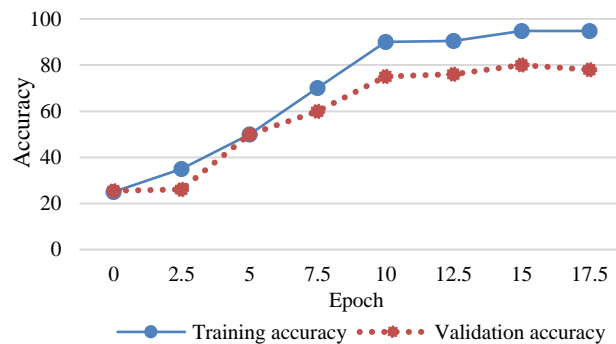


Figure 3. CNN-BiLSTM hybrid model accuracy

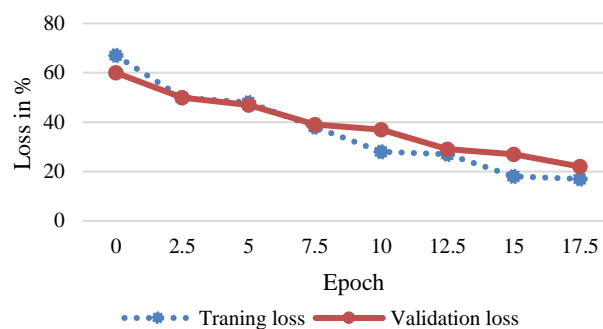


Figure 4. CNN-BiLSTM hybrid model loss graph

4. CONCLUSION

This paper proposed a CNN-BiLSTM model based on word embedding for Amharic text event classification. Specifically, in this work, different machine and deep learning models have been experimented with, the proposed hybrid CNN and BiLSTM model with word2vec word embedding performed better for Amharic text event classification compared to the SVM, RF, and NB. Because CNN extracts local features of a text and BiLSTM extracts global or contextual and semantic of a word using forward and backward layers, this prevents gradient disappearance and gradient explosion. The BiLSTM solves the problem that appears in LSTM which learns only the current word information and, CNN with BiLSTM through the fully connected layer. Finally, the hybrid model extracted local and global features with the context of a word to capture the dependency of words in a document. This model scored an accuracy of 94.8% for text event classification. In future work, we recommend the researchers investigate other variances of deep learning techniques such as transfer learning with contextual embedding to validate and confirm the result obtained in this study.




REFERENCES

- [1] M. V. Subbarao, K. Venkatarao, S. Chittineni, and S. Kompella, "Utilizing deep learning, feature ranking, and selection strategies to classify diverse information technology ticketing data effectively," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 4, pp. 1985–1994, Dec. 2023, doi: 10.11591/ijai.v12.i4.pp1985-1994.
- [2] F. Gholami, Z. Rahmati, A. Mofidi, and M. Abbaszadeh, "On enhancement of text classification and analysis of text emotions using graph machine learning and ensemble learning methods on non-English datasets," *Algorithms*, vol. 16, no. 10, Oct. 2023, doi: 10.3390/a16100470.
- [3] J. Liu, Y. Chen, K. Liu, W. Bi, and X. Liu, "Event extraction as machine reading comprehension," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, pp. 1641–1651, doi: 10.18653/v1/2020.emnlp-main.128.
- [4] J. Li and C. Wu, "Deep learning and text mining: classifying and extracting key information from construction accident narratives," *Applied Sciences*, vol. 13, no. 19, p. 10599, Sep. 2023, doi: 10.3390/app131910599.
- [5] Y. I. Alzoubi, A. E. Topcu, and A. E. Erkaya, "Machine learning-based text classification comparison: Turkish language context," *Applied Sciences*, vol. 13, no. 16, p. 9428, Aug. 2023, doi: 10.3390/app13169428.
- [6] D. Ali, M. M. S. Missen, and M. Husnain, "Multiclass event classification from text," *Scientific Programming*, vol. 2021, pp. 1–15, Jan. 2021, doi: 10.1155/2021/6660651.
- [7] Z. Subecz, "Event detection and classification in Hungarian natural texts," *European Scientific Journal ESJ*, vol. 15, no. 21, pp. 411–422, Jul. 2019, doi: 10.19044/esj.2019.v15n21p411.
- [8] S. Pandey, A. K. Srivastava, and B. G. Amidan, "A real-time event detection, classification and localization using synchrophasor data," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4421–4431, Nov. 2020, doi: 10.1109/TPWRS.2020.2986019.
- [9] N. Günemann-Gholizadeh, "Machine learning methods for detecting rare events in temporal data," M.S. thesis, Technische Universität München, 2018.
- [10] R. Khairunnas, J. A. Pagua, G. Fitriya, and Y. Ruldeviyani, "User sentiment dynamics in social media: a comparative analysis of X and Threads," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 1, pp. 447–456, Feb. 2025, doi: 10.11591/ijai.v14.i1.pp447-456.
- [11] B. Abera and B. Abera Hordofa, "Event extraction and representation model from news articles," *Article in International Journal of Innovations in Engineering and Technology*, vol. 16, no. 3, pp. 1–8, 2020, doi: 10.21172/ijiet.163.01.
- [12] E. Tadesse, R. T. Aga, and K. Qaqqabaa, "Event extraction from unstructured Amharic text," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020, pp. 2103–2109.
- [13] C. Spampinato *et al.*, "A rule-based event detection system for real-life underwater domain," *Machine Vision and Applications*, vol. 25, no. 1, pp. 99–117, May 2014, doi: 10.1007/s00138-013-0509-x.
- [14] A. Andualem and T. Tegegne, "Design event extraction model from Amharic texts using deep learning approach," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 411 LNICST, Springer International Publishing, 2022, pp. 424–434, doi: 10.1007/978-3-030-93709-6_28.
- [15] X. Liang, D. Cheng, F. Yang, Y. Luo, W. Qian, and A. Zhou, "F-HMTC: detecting financial events for investment decisions based on neural hierarchical multi-label text classification," in *IJCAI International Joint Conference on Artificial Intelligence*, Jul. 2020, pp. 4490–4496, doi: 10.24963/ijcai.2020/619.




- [16] F. Jenhani, M. S. Gouider, and L. Ben Said, "A hybrid approach for drug abuse events extraction from Twitter," *Procedia Computer Science*, vol. 96, pp. 1032–1040, 2016, doi: 10.1016/j.procs.2016.08.121.
- [17] H. L. Trieu, T. T. Tran, K. N. A. Duong, A. Nguyen, M. Miwa, and S. Ananiadou, "DeepEventMine: end-to-end neural nested event extraction from biomedical texts," *Bioinformatics*, vol. 36, no. 19, pp. 4910–4917, Jun. 2020, doi: 10.1093/bioinformatics/btaa540.
- [18] N. A. Mohammed, M. H. Abed, and A. T. Albu-Salih, "Convolutional neural network for color images classification," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 3, pp. 1343–1349, Jun. 2022, doi: 10.11591/eei.v11i3.3730.
- [19] V. Q. Nguyen, T. N. Anh, and H. J. Yang, "Real-time event detection using recurrent neural network in social sensors," *International Journal of Distributed Sensor Networks*, vol. 15, no. 6, Jun. 2019, doi: 10.1177/1550147719856492.
- [20] Y. Li, X. Wang, and P. Xu, "Chinese text classification model based on deep learning," *Future Internet*, vol. 10, no. 11, p. 113, Nov. 2018, doi: 10.3390/fi10110113.
- [21] X. Qin *et al.*, "Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews," *Journal of Clinical Epidemiology*, vol. 133, pp. 121–129, May 2021, doi: 10.1016/j.jclinepi.2021.01.010.
- [22] B. Barik, U. K. Sikdar, and B. Gambäck, "NTNU at SemEval-2018 Task 7: classifier ensembling for semantic relation identification and classification in scientific papers," in *Proceedings of the 12th International Workshop on Semantic Evaluation*, 2018, pp. 858–862, doi: 10.18653/v1/S18-1138.
- [23] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, "Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya," *Information*, vol. 12, no. 2, Jan. 2021, doi: 10.3390/info12020052.
- [24] L. S. Al-homed, K. M. Jambi, and H. M. Al-Barhamtoshy, "A deep learning approach for Arabic manuscripts classification," *Sensors*, vol. 23, no. 19, Sep. 2023, doi: 10.3390/s23198133.
- [25] T. Sabri, S. Bahassine, O. El Beggat, and M. Kissi, "An improved Arabic text classification method using word embedding," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 721–731, Feb. 2024, doi: 10.11591/ijece.v14i1.pp721-731.
- [26] W. B. Demilie, A. O. Salau, and K. K. Ravulakollu, "Evaluation of part of speech tagger approaches for the Amharic language: a review," in *Proceedings of the 2022 9th International Conference on Computing for Sustainable Global Development, INDIACom 2022*, Mar. 2022, pp. 569–574, doi: 10.23919/INDIACom54597.2022.9763213.
- [27] B. T. Abeje, A. O. Salau, H. A. Ebabu, and A. M. Ayalew, "Comparative analysis of deep learning models for aspect level Amharic news sentiment analysis," in *2022 International Conference on Decision Aid Sciences and Applications, DASA 2022*, Mar. 2022, pp. 1628–1633, doi: 10.1109/DASA54658.2022.9765172.

BIOGRAPHIES OF AUTHORS






Amogne Andualem Ayalew    holds received a Master of Science degree in information technology from Bahir Dar University, Ethiopia 2021. He received his B.Sc., in information technology from Debre Markos University, Ethiopia in 2017. His research interests include natural language processing, image processing, machine learning, data mining, and network security. He can be contacted at email: amogneandualem@gmail.com.






Melaku Lake Tegegne    holds an MSc in information technology from Bahir Dar University, Ethiopia in 2021. He received his BSc in information technology from Debre Markos University, Ethiopia in 2017. His research interests include natural language processing, computer vision, health informatics, bioinformatics, machine learning, nano technology, and quantum computing. He can be contacted at email: mellmik125@gmail.com.






Bommy Manivannan    is currently working as an assistant professor in the Department of Computer Science & Engineering at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. She can be contacted at: bommy@mits.ac.in.






Tamarasi Suresh    is heading the Department of Information Technology at St.Peter's Institute of Higher Education and Research. Her broad areas of research interest are machine learning, software engineering, artificial intelligence, and cloud computing. Her journey in SPIHER started 2 years ago and it is been a wonderful environment to nurture her academic capabilities. She can be contacted at email: tamarasisu@gmail.com.






Napa Komal Kumar    is currently working as an assistant professor in the Department of Computer Science & Engineering (Data Science) at Madanapalle Institute of Technology & Science, Madanapalle. His research interests include machine learning, data mining, and cloud computing. He can be contacted at email: komalkumarna@gmail.com.






Battula Krishna Prasad    is currently working as an assistant professor in the Department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India. His research interests include machine learning, image processing, and serverless computing. He can be contacted at email: bkrishnaprasad@kluniversity.in.



Tsehay Admassu Assegie    holds a Master of Science degree in computer science from Andhra University, India 2016. He also received his B.Sc., in computer science from Dilla University, Ethiopia in 2013. Currently, he is pursuing Ph.D. in the Department of Electronic and Electrical Engineering, at Kyungpook National University, Daegu, Republic of Korea. His research interest includes medical image processing and the application of artificial intelligence in the healthcare. He has published over 61 scholarly papers in reputed international journals and international conferences. He can be contacted at email: tsehayadmassu2006@gmail.com.



Ayodeji Olalekan Salau    received a B.Eng. in electrical/computer engineering from the Federal University of Technology, Minna, Nigeria. He received his M.Sc. and Ph.D. degrees from the Obafemi Awolowo University, Ile-Ife, Nigeria. Dr. Salau is an indefatigable luminary in the fields of research and academia, whose profound expertise encompasses a multifaceted spectrum of subjects, including "Research in the fields of computer vision, image processing, signal processing, machine learning, control systems engineering, and power systems technology." With an unquenchable thirst for knowledge and an unwavering dedication to unraveling complex problems, Dr. Salau has etched an indelible mark in the domain of computer vision and image processing, meriting accolades for contributions to the field. Dr. Salau is a recipient of the Quarterly Franklin Membership with ID number CR32878 given by the Editorial Board of London Journals Press in 2020 for top-quality research output. Dr. Salau's research paper was awarded the best paper of the year 2019 in Cogent Engineering. Currently, Dr. Salau works at Afe Babalola University as an associate professor in the Department of Electrical/Electronics and Computer Engineering. He can be contacted at email: ayodejisalau98@gmail.com.