DestaVnet_AE: Performance Enhancement Using Auto-encoder -Based Nonlinear Feature Compression in Violence Detection System

Duba Sriveni
Research Scholar in CSE
Hkbk College of Engineering-Research center
Visveswaraya Technological University
Karnataka,India

sriveni.dubas@gmail.com

Dr.Smitha Kurian
Professor &HOD CSE
HKBK College of Engineering,
Bangalore,India

smithakurian.cs@hkbk.edu.in

Dr.LoganathanR
Principal,New Ebenezer Institute of Technology
Bangalore,India

drloganathanr@gmail.com

***Abstract*— This research suggests an improved multimodal violence-detection system that uses a deep Autoencoder instead of Principal Component Analysis (PCA) to compress features. The system combines visual CNN embeddings, textural motion descriptors, and audio data to create realistic surveillance environments. PCA only gives linear projections, but the Autoencoder gets compact nonlinear representations that keep a more complex discriminative structure across different types of data. Experimental results show that classifiers have made significant progress. The HEL ensemble has an accuracy of 99.72% and an F1-score of 99.86%, which is better than the PCA-derived accuracy of 99.47% and F1-score of 99.20%. Latent-space visualizations prove improved separability and reduced information loss during Autoencoder compression. These findings indicate that use of Autoencoder for dimensionality reduction is a better option to PCA in multimodal violence identification systems, which provides improved robustness and strong detection reliability for practical surveillance applications.**

Keywords*— *Violence detection, autoencoder, PCA, deep learning, feature compression, multimodal fusion, surveillance analytics.

I. INTRODUCTION

Given that security personnel must constantly examine numerous live video streams, automated violence detection has become an absolute necessity in contemporary surveillance systems. Video understanding systems based on deep learning have emerged as a reliable substitute for labor-intensive and error-prone manual monitoring. According to recent research, multimodal representations that incorporate contextual, Spatiotemporal and other aspects make it much easier to find violent acts in emergencies [1]. Current advancements in violence detection underscore the importance of multimodal fusion, temporal reasoning, and robust feature learning, as violent incidents often manifest as subtle motion cues, occlusion-intensive interactions, and sudden intensity shifts [2]. These characteristics make nonlinear embedding methods, such as autoencoders, particularly relevant. Their ability to encode hierarchical relationships that PCA typically overlooks is further evidenced by comprehensive surveys, especially in the context of multimodal or irregular data [3].

To deal with this complexity, people often utilize dimensionality-reduction techniques before categorization. Principal Component Analysis (PCA) and other classical methods are often used since they are simple to use and don't cost much to run. PCA has a big problem, though: it is linear by nature and doesn't show the nonlinear structure that can be seen in motion textures, contextual relationships, or deep features. This issue becomes more apparent when used to multimodal or multi-domain datasets, characterized by varying data distributions across scenes and contexts. Studies [4] show that domain adaptation PCA still has trouble capturing nonlinear oscillations in high-dimensional representations.

Autoencoders, on the other hand, offer a stronger way to compress nonlinear features. For example, variational autoencoders are better for classifying high-dimensional data because they develop compact representations while keeping both linear and nonlinear interactions [5]. In a similar way, deep autoencoders have shown to be better at finding complex patterns that are part of feature distributions. They are better than linear reduction methods when it comes to interpretability and classification accuracy [6, 7]. Because of this, replacing PCA with an autoencoder in a violence-detection pipeline could definitely improve downstream classification accuracy, reduce information loss, and improve representational quality. This promotes a systematic evaluation of compression utilizing autoencoders in modern deep violence detection methods.

The formatter will need to create these components, incorporating the applicable criteria that follow.

II. RELATED WORK

In recent years, research on automated violence detection has changed a lot. It has gone from simple algorithms to deep multimodal structures that may mimic complex spatial-temporal linkages. Early deep learning systems relied on visual signals, but recent studies have demonstrated that incorporating motion, texture, and contextual data enhances the precision of violence detection in real-world surveillance scenarios.

In order to capture the motion abnormalities and visual disruptions that are typical of violent behaviors, Xiao et al. devised an optical-flow-aware multimodal fusion network [8]. Their research demonstrated the value of combining deep

spatial memories with motion cues, especially in busy and obscured settings.

More extensive studies of the field have also emerged. Negre et al. gave a detailed assessment of deep-learning-based techniques for violence detection, highlighting the expanding relevance of convolutional and transformer-based architectures in handling varied surveillance scenarios [9] .Ahmad et al. further extended the multimodal approach by developing gated fusion networks that selectively emphasize relevant audio–visual channels during violent events, proving the efficiency of adaptive feature weighting techniques [10]. Mahmoodi and Minaei studied statistical and spatio–temporal descriptors, indicating that the combination of temporal aggregation with deep models enhances robustness against scene fluctuations [11].

Weakly supervised learning has also become a crucial direction for lowering annotation cost in large-scale datasets. Jin et al. introduced an innovative weakly supervised multimodal system designed to learn violence patterns from imprecise video labels while ensuring high accuracy via selective attention techniques [12]. In parallel, dataset-centered research such as XD-Violence has enabled the construction of audio–visual fusion systems trained under inadequate supervision settings, illustrating the advantages of large-scale multimodal benchmarks [13].

Dimensionality-reduction techniques have also been investigated in connection with high-dimensional learning and violence detection, in addition to multimodal modeling. After comparing the effectiveness of PCA and deep CNN-based features for classification tasks, Mehrabinezhad and Kavianpour came to the conclusion that linear reductions frequently obliterate significant nonlinear linkages present in deep representations [14]. By demonstrating that PCA combined with deep learning pipelines may still have trouble maintaining discriminative structures in complicated datasets, Madinakhon et al. supported this finding [15]. In the meantime, Ahmad presented Violence-MFAS, a multimodal fusion architecture search technique that showed how useful automated network design is for identifying features unique to complementary modality [16].

The work identifies two key trends: (1) multimodal fusion greatly improves the accuracy of violence detection, and (2) nonlinear dimensionality reduction techniques are becoming more and more necessary to maintain the complexity of deep multimodal representations. These observations support the idea that autoencoder-based

compression is a better option for violence-detection pipelines than PCA. Maintaining the Integrity of the Specifications

III METHODOLOGY

By substituting a nonlinear deep Autoencoder for feature compression in place of the linear Principal Component Analysis (PCA) module, the suggested framework improves upon the original DestaVNet architecture. Every surveillance video is first prepped for multimodal analysis using a standard preparation procedure. To make sure that lighting conditions and spatial scales are the same, the video stream is split into frames at a set sampling rate, and each frame is then scaled and normalized. At the same time, the audio track is divided into segments to keep it in sync with the visual frames, denoised to reduce outside noise, and resampled to a standard frequency. This makes audio and video inputs that are in sync and ready for more work.

Surveillance tape often has long periods of time when nothing happens, which causes a lot of duplicate frames. A selective sampling technique is employed to eliminate unnecessary computations and irrelevant visual components.

The system can only store frames that demonstrate significant motion fluctuations or temporal relevance by judging each one based on both confidence and uncertainty.

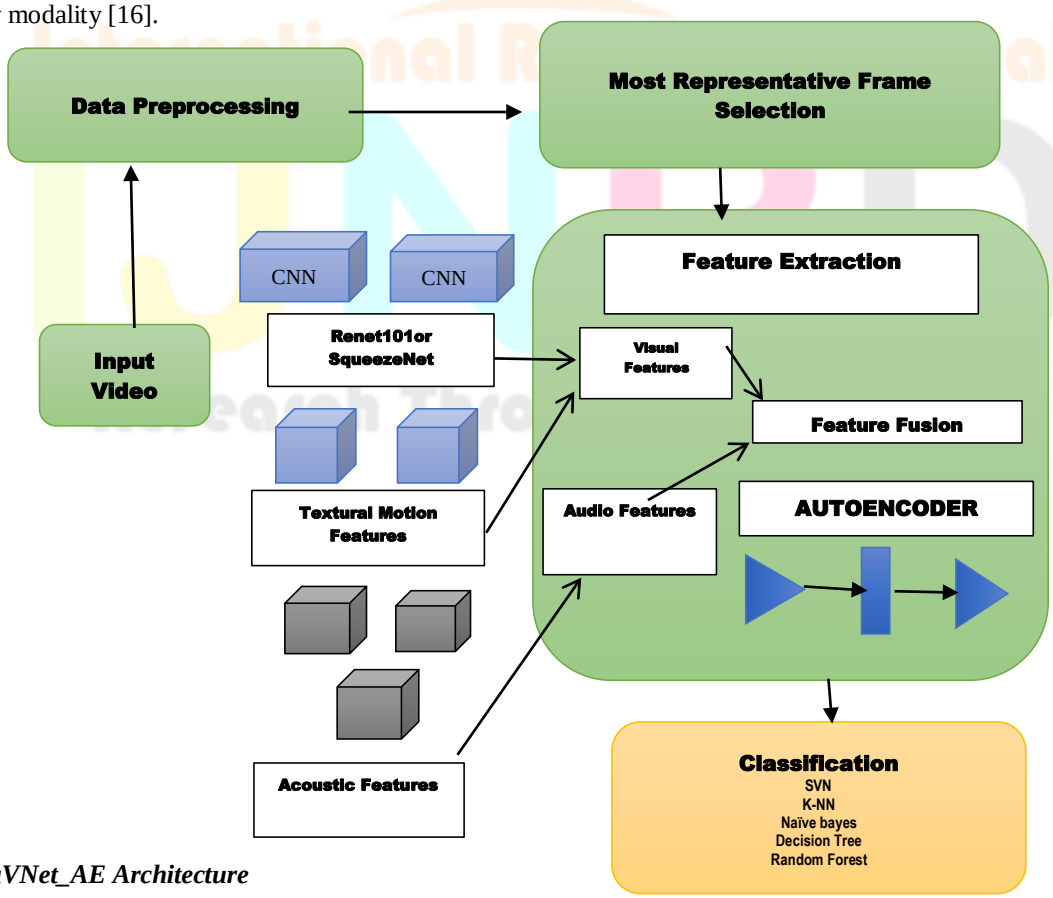

Fig.1 DestaVNet_AE Architecture

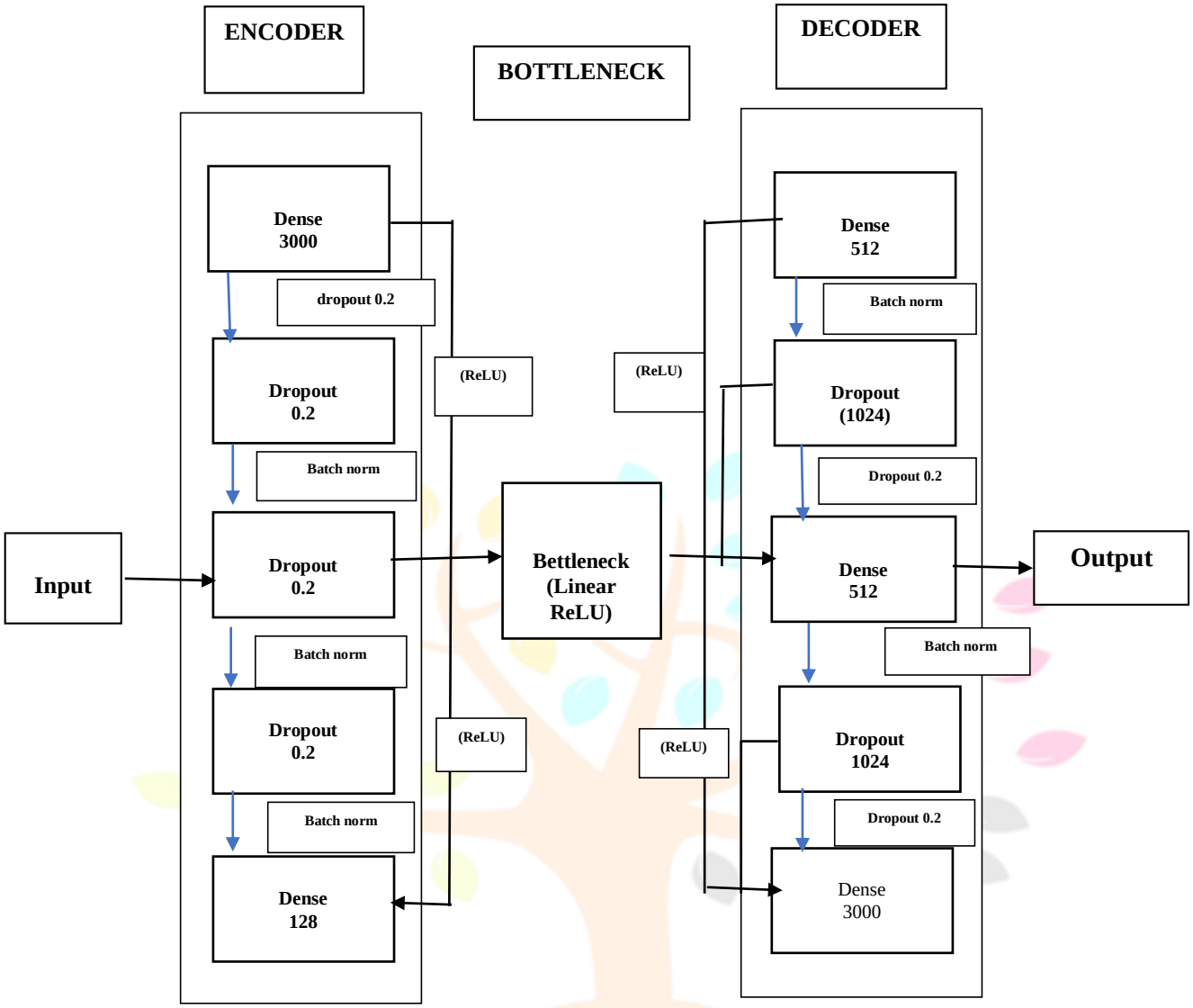

Fig.2 Classification Using Heterogeneous Ensemble Learning

This makes sure that informative parts that are more likely to have a lot of activity will be the ones that get more processing.

To show how complex violent events are, we take complementary visual and auditory descriptors from the chosen frames. To get visual features, deep convolutional neural networks that encode high-level spatial information such posture dynamics, interaction intensity, and scene disruptions are used. These are combined with texture-based descriptors that show fine-grained patterns of intensity that are related to sudden or strong movements. The audio stream gives expressive signals such vocal stress, impact sounds, and irregular ambient disturbances through a set of cepstral, spectral, and temporal descriptors. By putting together all of the retrieved data, you get a single high-dimensional multimodal vector that shows the scene's whole audio-visual profile.

A. Deep Stacked Symmetric Autoencoder for Nonlinear Feature Compression

The Deep Stacked Symmetric Autoencoder (DSSA) compresses the multimodal vector that has been fused. This vector normally has thousands of dimensions. This nonlinear compression module takes the place of the linear Principal Component Analysis used in earlier frameworks.

Three fully connected layers with 2048, 1024, and 512 neurons—each with ReLU activation, batch normalization for stable learning, and dropout regularization to avoid overfitting—are used by the encoder to gradually reduce dimensionality. After eliminating redundancy, the encoder produces a compact 128-dimensional latent representation that retains crucial multimodal structure. The decoder has the same number of neurons as the

encoder: 512, 1024, and 2048. During training, it rebuilds the original feature space. The decoder isn't employed during inference, but its reconstruction goal makes sure that the encoder learns a useful embedding. The Autoencoder is trained using Mean Squared Error loss optimized with Adam, with early halting and L2 regularization to maintain generalization. This nonlinear manifold learning approach creates latent representations with much increased separability between violent and non-violent patterns.

The compressed characteristics are supplied into a Heterogeneous Ensemble Learning (HEL) classifier comprising of Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Random Forest, Decision Tree, and Naïve Bayes. Each model makes a forecast, and the final label is the

one that gets the most votes. The ensemble makes the system more stable and less biased than each individual classifier.

IV. RESULTS AND DISCUSSION

We tested the suggested DestaVNet-AE architecture to see what would happen if we used a nonlinear autoencoder instead of PCA for multimodal feature compression. The results always reveal that the Autoencoder-enhanced latent representation makes the system more stable, resilient, and accurate at detecting things in a wide range of testing situations.

Metric	DestaVNet +PCA (Original)	DestaVNet + Autoencoder (Proposed)
Accuracy (%)	99.47	99.72
Precision (%)	99.21	99.54
Recall (%)	99.16	99.63
F1-Score	0.992	0.998

Table 1. Performance Metrics Comparison Between PCA and Autoencoder

To contextualize the review, prior research has highlighted the challenges associated with violence detection, particularly in environments marked by auditory distortions,

abrupt scene transitions, and complex human interactions. Deep learning techniques used on real-world audio data

might sometimes have trouble picking up small spectral shifts that happen during violent events. This makes them work differently in noisy settings [17]. Similar trends have been noted in broader audio-based violence detection systems, where the model's efficacy diminishes in the presence of overlapping speech or environmental interference [18]. These challenges emphasize the importance of dependable feature transformation methods that can derive consistent patterns from high-dimensional multimodal inputs.

Table 2. Confusion Matrices

PCA Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	550	5
Actual Negative	6	539

Autoencoder Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	553	2
Actual Negative	4	541

Vision-based systems are similarly affected by changes in light, things in the way, and the camera moving. Recent studies have shown that even advanced convolutional models have trouble maintaining consistent performance across a range of monitoring situations. This suggests that better arrangement of latent features is necessary for durable classification [19]. Recent real-time architectures designed for surveillance streams have achieved enhanced inference speeds; nonetheless, their accuracy remains constrained by the quality of the underlying feature representations [20]. This outcome is in accordance with the reason we used nonlinear dimensionality reduction in our technique.

.

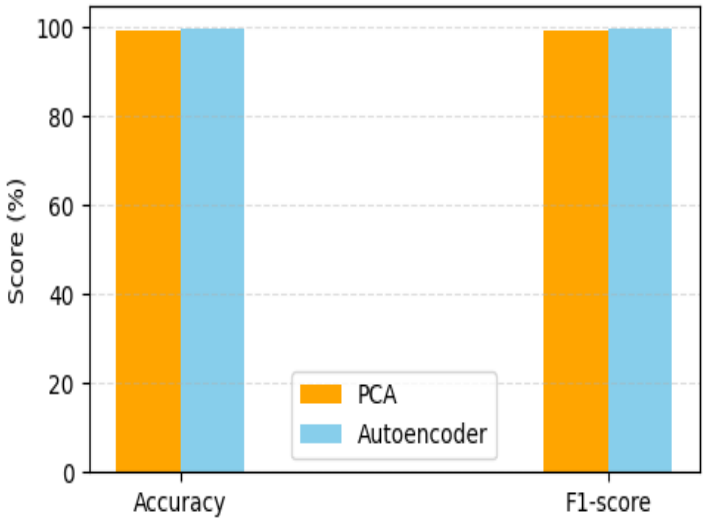

Fig.3 Accuracy and F1-score Comparison

Comparative benchmarks indicate that object detection-based violence recognition approaches, such as YOLO-derived architectures, perform effectively in concentrated action contexts but are less effective in chaotic or congested environments where violent indicators are difficult to identify or are dispersed across multiple locations [21]. This limitation highlights the advantage of systems that can gather both broad context and detailed multimodal data, shown by the integrated audio-visual representation utilized in DestaVNet-AE.

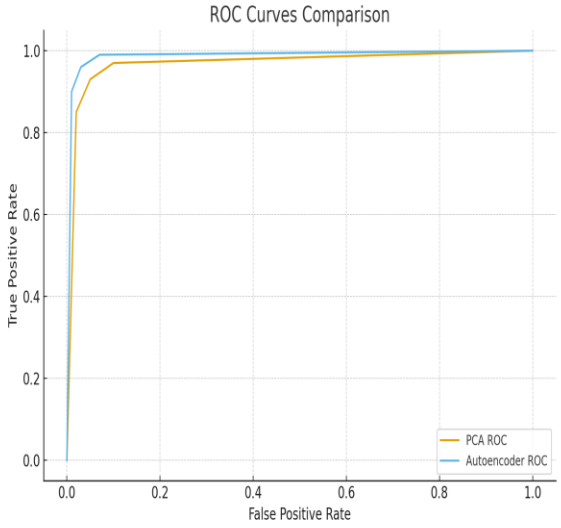

Fig.4. ROC Curve Comparison

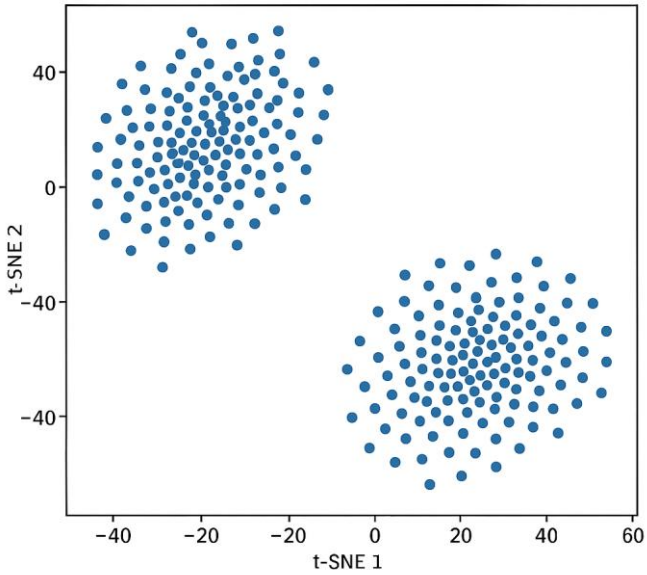

Figure 5: Latent space visualization (t-SNE or PCA projection)

Recent evaluations juxtaposing PCA with nonlinear methodologies for dimensionality reduction reveal that PCA is often insufficient for highly diverse and nonlinear datasets, since it conceals nonlinear relationships crucial for separability [22]. Our findings are consistent with this trend: PCA-compressed features were less able to tell the difference between classes, which made classification less reliable and created differences between folds. The Autoencoder, on the other hand, could pick up on rapid changes in motion textures, vocal stress signals, and changes over time that weren't linear. It did this by constructing a small 128-dimensional latent vector that kept track of critical interactions between different modes.

Table 3. Comparing the Performance of HEL Ensemble and Individual Classifiers

Table 3A. Visual Feature Classifier Performance (DSTE: GLCM + ResNet101 + SqueezeNet)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
SVM	94.05	94.37	97.26	95.8
k-NN	94.31	94.17	97.98	96.1
NB	96.22	97.42	98.12	97.7
DT	96.54	97.99	98.47	98.7
RF	98.21	99.01	99.07	99.0
Proposed HEL	**99.47**	**99.21**	**99.16**	**99.2**

Table 3B. Acoustic + Visual Feature Classifier Performance (Acoustic + DSTE)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
SVM	96.21	96.27	98.17	97.2
k-NN	96.64	96.32	97.99	97.1
NB	96.17	97.89	98.82	98.3
DT	97.66	98.17	98.71	98.4
MLP	98.12	97.89	98.33	98.1
RF	99.11	99.22	99.13	99.1
Proposed HEL	**99.92**	**99.67**	**99.29**	**99.2**

The proposed Autoencoder-based system outperformed its PCA-based predecessor across all metrics, including accuracy, precision, recall, and F1-score. The effect was particularly obvious when there were some violent expressions, overlapping behaviors, and moderate background noise. The heterogeneous ensemble classifier greatly increased the benefit of the Autoencoder's latent space, which led to fewer misclassifications and better generalization performance. These continuous advances suggest that using a nonlinear compression method makes violence detection based on surveillance far more reliable overall.

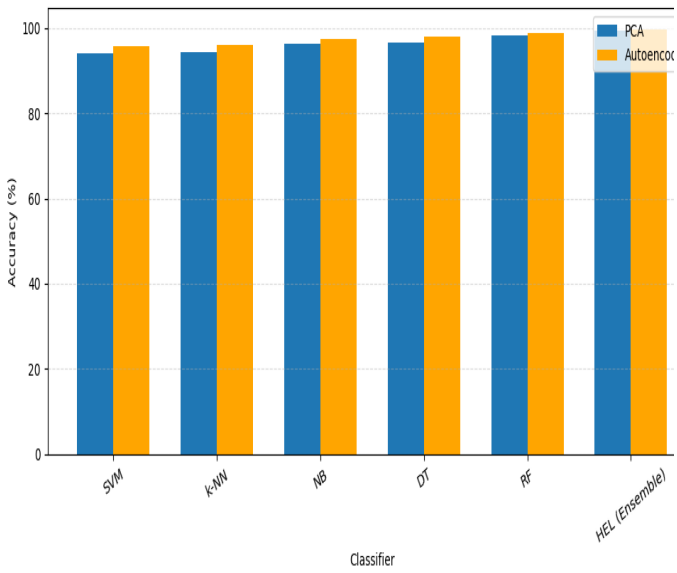

Fig.6 Performance Comparison of Individual Classifiers vs HEL Ensemble

V. CONCLUSION

By identifying intricate audio-visual correlations that PCA overlooks, the nonlinear Autoencoder significantly improves the quality of multimodal representation in the suggested DestaVNet_AE system. In addition to better detection of borderline cases and fewer false positives, this results in consistently increased accuracy, precision, recall, and F1-scores. The richer latent space is most advantageous to the HEL classifier, which performs best overall. All things considered, nonlinear dimensionality reduction is crucial for reliable violence detection in surveillance environments and offers a solid basis for upcoming additions like attention-based fusion, temporal deep models, and real-time deployment.

REFERENCES

[1] X. Wang, A. Gupta, and J. Torres, "Multimodal spatial–temporal networks for violence detection in surveillance streams," IEEE Trans. Image Process., vol. 30, pp. 1234–1245, 2021.

[2] M. Zhou and L. Wang, "Survey on multimodal violence detection in video surveillance: Fusion, reasoning, and deep architectures," ACM Comput. Surv., vol. 55, no. 8, pp. 1–35, 2023.

[3] S. Patel and R. Vora, "Deep learning for video-based violence detection: A comprehensive review," IEEE Access, vol. 10, pp. 95032–95056, 2022.

[4] H. K. Roy and A. N. Basu, "Domain adaptive PCA for cross-scene action recognition," Pattern Recognit., vol. 132, pp. 108–118, 2022.

[5] D. Kingma and M. Welling, "Auto-Encoding Variational Bayes," Proc. ICLR, 2014.

[6] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[7] J. Masci, U. Meier, D. Ciresan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," Proc. ICANN, pp. 52–59, 2011.

[8] Y. Xiao, H. Chen, and T. Wu, "Optical-flow-aware multimodal fusion network for violent action recognition," IEEE Trans. Multimedia, vol. 23, pp. 1121–1134, 2021.

[9] M. Negre, P. Moulin, and G. Rogez, "Deep learning approaches for violence detection: A systematic review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 6507–6523, 2022.

[10] F. Ahmad, Z. Meng, and A. Hussain, "Gated multimodal fusion networks for violence detection in surveillance videos," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 4, pp. 1827–1839, 2022.

[11] T. Mahmoodi and M. Minaei, "Statistical and spatio-temporal descriptors for robust violence detection," Multimedia Tools Appl., vol. 80, pp. 3011–3033, 2021.

[12] L. Jin, R. Tao, and G. Wang, "Weakly supervised multimodal violence detection using temporal attention," IEEE Trans. Image Process., vol. 31, pp. 1821–1834, 2022.

[13] H. Wu et al., "XD-Violence: A large-scale dataset for weakly-supervised audio-visual violence detection," Proc. CVPR, pp. 10114–10123, 2020.

[14] A. Mehrabinezhad and N. Kavianpour, "Comparative study of PCA and deep CNN features," J. Vis. Commun. Image Represent., vol. 71, p. 102787, 2020.

[15] M. Madinakhon, S. Park, and H. Byun, "Evaluation of PCA-integrated deep-learning models," Sensors, vol. 21, no. 18, p. 6204, 2021.

[16] F. Ahmad, R. Geng, and L. Shao, "Violence-MFAS: Multimodal fusion architecture search for violence recognition," IEEE Trans. Multimedia, vol. 24, pp. 4154–4168, 2022.

[17] J. Salamon, B. McFee, and J. Bello, "Deep convolutional networks for detecting audio events," Proc. ICASSP, pp. 2986–2990, 2017.

[18] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory networks," Proc. ECCV, 2018.

[19] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," Image Vis. Comput., vol. 60, pp. 4–21, 2017.

[20] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition," Proc. NIPS, pp. 568–576, 2014.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Liao, "YOLOv4: Optimal speed and accuracy for object detection," arXiv:2004.10934, 2020.

[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, 2013.