SenSeqNet: A Deep Learning Framework for Cellular Senescence Detection from Protein Sequences

Hanli Jiang[3#], Li Lin [1, 2#], Dongliang Deng[4#], Jianyu Ren[5], Xin Yang[6], Siyi Liu[1, 2],Lubin Liu[1, 2*]

[1]Department of Obstetrics and Gynecology, Women and Children's Hospital of Chongqing Medical University, Chongqing 401147, China.

[2]Department of Obstetrics and Gynecology, Chongqing Health Center for Women and Children, Chongqing 401147, China.

[3]Department of Computer Science, University of Toronto, Toronto M5S1A1, Canada.

[4]Department of Oncology, Chongqing Traditional Chinese Medicine Hospital, Chongqing 400021, China.

[5]Department of Computer Science, University of Wisconsin - Madison, Madison, WI 53706 , USA.

[6]Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA.

#These authors contributed equally to this work.

* Correspondence to Lubin Liu: No.120 Longshan Road, Yubei District, Women and Children's Hospital of Chongqing Medical University, Chongqing, 401147, China.

Email: liulubin@hospital.cqmu.edu.cn

Abstract

Cellular senescence, characterized by the irreversible cessation of division in normally proliferating cells due to various stressors, presents a significant challenge in the treatment of age-related diseases. Understanding and accurately detecting cellular senescence is crucial for identifying potential therapeutic targets. However, traditional wet lab assays for detecting cellular senescence are time-consuming and labor-intensive, limiting research and drug development efficiency. There is an urgent need for computational tools allowing swift and accurate detection of cellular senescence from protein sequences. We propose SenSeqNet, a novel deep learning framework for detecting cellular senescence directly from protein sequences. The framework begins with feature extraction using the Evolutionarily Scaled Model (ESM-2), a state-of-the-art protein language model that captures evolutionary information and complex sequence patterns. The extracted embeddings are then passed through a hybrid architecture consisting of long short-term memory (LSTM) networks and convolutional neural networks (CNNs) to further refine and learn from the embedded information. SenSeqNet achieved a final accuracy of 83.55% on independent testing, surpassing

various machine learning and deep learning architectures. This performance underscoring the robustness and effectiveness of SenSeqNet for detecting cellular senescence from protein sequences. These results provide a solid foundation for future research on aging and age-related therapeutics.

Key words： Cellular senescence, Protein language model, Deep learning

1. Introduction

Aging is a complex biological process marked by the gradual accumulation of cellular and molecular damage over time, leading to diminished physiological function and the onset of age-related diseases[1]. Understanding the underlying mechanisms of cellular senescence is essential for developing interventions that could mitigate these effects[2]. Central to this pursuit is the accurate detection and classification of aging cells, which offers crucial insights into the molecular pathways driving the aging process. Recent advances in high-throughput sequencing and proteomics have generated an unprecedented volume of biological data, presenting new opportunities to explore aging at the molecular level. Protein sequences, in particular, are of significant interest because they play a direct role in nearly all cellular processes. Analyzing these sequences to identify signs of aging could reveal key biomarkers and therapeutic targets, paving the way for novel interventions in age-related conditions.

Traditional methods for studying protein sequences have often relied on manual analysis and rule-based approaches, which are limited in their capacity to uncover the complex patterns associated with aging[3]. However, the emergence of machine learning (ML) and deep learning (DL) has revolutionized the analysis of biological data, enabling the automatic extraction of complex patterns and insights that were previously unattainable with traditional methods[4,5]. In particular, deep learning models, with their ability to automatically learn hierarchical representations from raw data, have demonstrated remarkable success in various biomedical applications, including protein structure prediction, drug discovery, and disease diagnosis[6-8].

Despite significant advancements in deep learning, the application of these techniques to detect cell aging using protein sequences remains relatively scarce. Recently, the integration of multiple deep learning models to leverage their respective advantages has demonstrated significant success[9]. To further explore the application of integrated deep learning models, we developed SenSeqNet (Fig. 1) , a novel deep learning approach that begins with ESM2 for feature extraction and then synergistically combines the strengths of long short-term memory (LSTM) networks and convolutional neural networks (CNNs)[10]. LSTM networks are adept at modeling sequential data like protein sequences, capturing long-term dependencies that are critical for understanding the temporal aspects of aging[11]. Meanwhile, CNNs excel at extracting spatial features[12], making them ideal for identifying complex patterns within the outputs generated by LSTM networks.

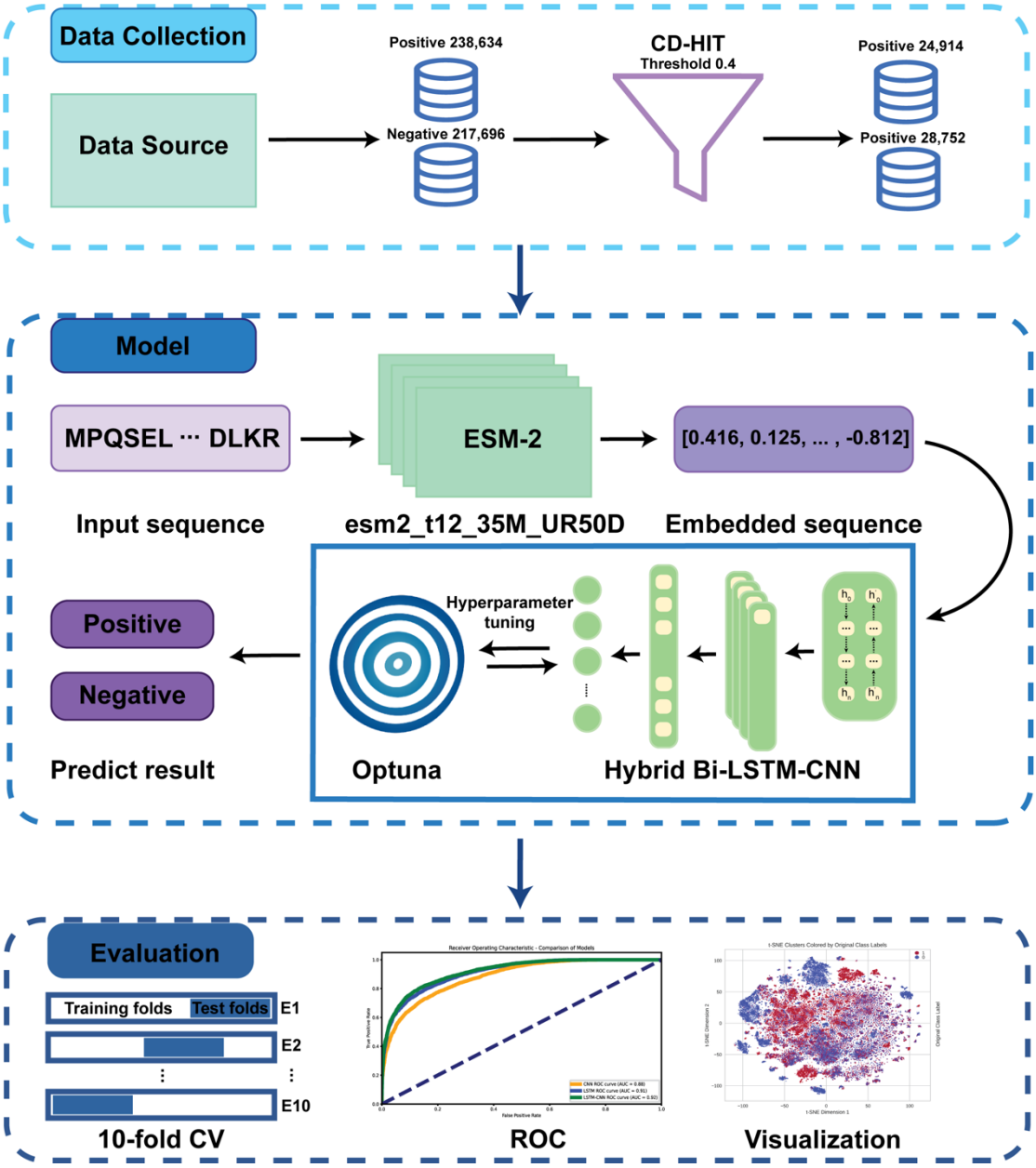

Fig. 1. The overall framework.

The next crucial step in sequence-based protein prediction tasks is choosing the optimal model architecture, with the representation of sequence data being a key factor in determining overall model performance[13]. We utilized the Evolutionarily Scaled Model (ESM)[14], a state-of-the-art transformer-based model renowned for its ability to capture deep evolutionary relationships in protein sequences. Specifically, we evaluated several variants of the ESM-2 model to generate embeddings for our task. To gain insights into the high-dimensional data structure, we conducted visualization analysis using the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm[15]. This allowed us to observe distinct clustering patterns within the protein sequence embeddings, highlighting the effectiveness of our embedding approach. We also compared the t-SNE visualizations between ESM-2 and other traditional feature extraction methods. ESM-2 outperforms

other approaches in producing more clearly defined and separable clusters. Through extensive experimentation, we evaluated multiple high-performance machine learning and deep learning models to identify the optimal architecture for classifying cell aging. We compared these models based on their ability to capture unique patterns in protein sequences related to cellular senescence. Ultimately, the SenSeqNet demonstrated superior performance from the independent testing, leveraging its ability to effectively process both sequential and spatial features. Furthermore, we conducted 10-fold cross-validation, demonstrating that SenSeqNet is a highly robust tool for this task. This study advances the application of deep learning in aging research and lays a foundation for identifying biomarkers and developing targeted therapies to promote healthy aging.

2. Materials and methods

2.1. Benchmark datasets

For the development of an effective prediction model, the dataset is generally composed of comparable positive and negative samples. In this study, cell senescence genes were downloaded from the CellAge: The Database of Cell Senescence Genes, where abnormal expression of these genes is associated with the induction of senescence. The database can be accessed at https://genomics.senescence.info/cells/index.html. Protein sequences related to cell senescence were downloaded from the Universal Protein Knowledgebase (UniProtKB)[16] and used as positive samples. In addition, protein sequences from housekeeping genes were selected as negative samples, resulting in a dataset comprising 238,634 positive and 217,696 negative protein sequences. To reduce redundancy and ensure high-quality data, we employed CD-HIT[17] with a threshold of 0.4, ultimately leaving 24,914 positive and 28,752 negative sequences. The dataset will be divided into a training set (80%) and a validation set (20%) to facilitate model development and performance evaluation.

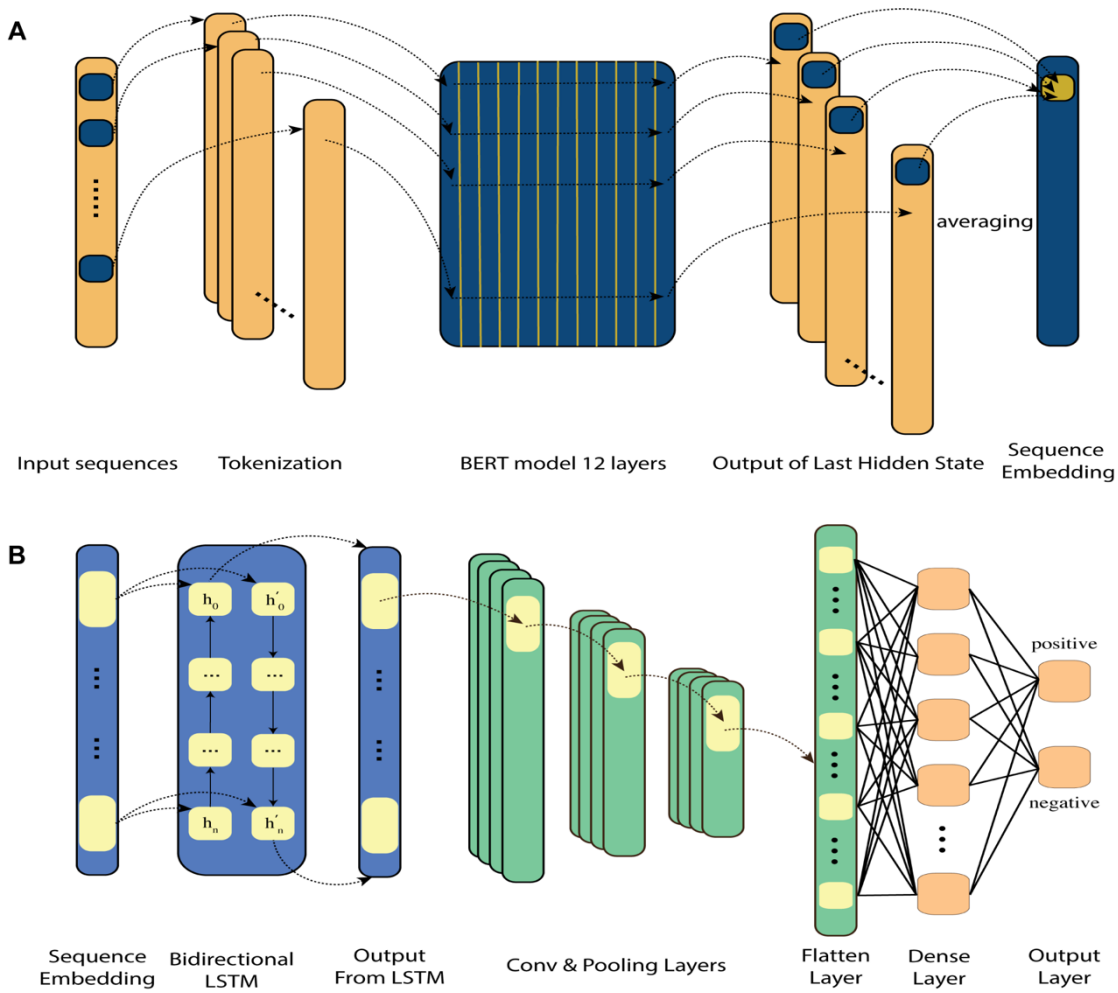

Fig 2. Sample Architecture of the SenSeqNet: A. Evolutionarily Scaled Model (ESM-2) architecture for feature extraction. B. Hybrid Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN) model.

2.2. SenSeqNet model structure

As illustrated in Figure 2, SenSeqNet consists with two main components: ESM-2 and the hybrid LSTM-CNN architecture. Beginning with the first component, effective representation of sequence data is an essential step in building a robust prediction model, especially when dealing with protein sequences[18]. Thus, we employed the ESM-2[14], known for its exceptional ability to capture deep evolutionary relationships in protein sequences (Fig. 2A). ESM-2 is trained on approximately 65 million unique protein sequences using a masked language modeling objective, which involves predicting the identity of randomly selected amino acids within a sequence based on their surrounding context[12]. This training enables ESM-2 to learn the dependencies between amino acids, allowing it to effectively capture both local and global contextual information.

The unsupervised nature of the ESM-2 model's training allows it to internalize sequence patterns across diverse proteins, effectively capturing structural information linked to these patterns[19]. ESM-2 is highly effective at processing and embedding from large-scale protein datasets, making it particularly suitable for sequence representation tasks. The model's performance across various

bioinformatics applications has been well-documented[20], demonstrating its ability to discern subtle patterns within sequences, which are important for accurate prediction tasks. There are multiple ESM-2 models, each differing in architectural configuration, where an increase in the number of transformer layers corresponds to a proportional increase in the number of model parameters. To identify the most effective variant for our dataset, we evaluated each variant's output and utilized the results to train two distinct models: a CNN to assess the spatial features and an LSTM to analyze the sequential dependencies within the results. By balancing the accuracy, computational efficiency, and scalability of these models, we determined the most effective ESM-2 variant for our dataset. The selected model variant was then used for embedding generation throughout our study. The detailed results of this evaluation will be discussed in the Results section.

Followed by ESM-2, the embedded sequences are input into the second main component, a hybrid LSTM-CNN architecture (Fig. 2B). In this setup, a Bidirectional LSTM network is first employed to process the sequence data, leveraging its capacity to model temporal dependencies in both directions, thus capturing comprehensive long-term relationships between amino acids in the protein sequences. These sequential features are then passed to the CNN, which focuses on identifying spatial patterns within the LSTM-extracted features. This combined approach enables the model to capture both temporal and spatial characteristics. This two-part architecture, combining the evolutionary context captured by ESM-2 with the temporal and spatial analysis capabilities of the LSTM-CNN, enables SenSeqNet to effectively identify crucial features in protein sequences. This comprehensive approach significantly enhances the model's ability to distinguish aging cells from non-aging cells, as demonstrated in the results.

2.3. Comparative models for benchmarking SenSeqNet

In this study, we explored several machine learning and deep learning architectures to develop a reliable model for detecting cell aging using protein sequences[18]. Given the lack of prior research in this specific domain of cell aging, our approach involved comprehensive experimentation with a range of models commonly employed in protein sequence prediction to understand their capabilities and limitations. The architectures evaluated include CNNs[21], Recurrent Neural Networks (RNNs)[22], LSTMs[11], and bidirectional LSTMs. Recognizing the strengths of both CNNs in extracting spatial features and bidirectional LSTMs in capturing sequential dependencies[23], we also experimented with hybrid models that combined these two architectures. The combined CNN and bidirectional LSTM model was designed to leverage the advantages of both, aiming to improve the accuracy and robustness of cell aging detection from protein sequences. Additionally, we constructed several machine learning models, including Bagging[24], Support Vector Machines (SVM)[25], Random Forest (RF)[26], and XGBoost[27], to provide a comparative understanding of their performance against deep learning approaches. Below, we present a comprehensive description of each deep learning model. Hyperparameters such as learning rate, dropout rate, batch size, and the number of layers were optimized for each model using Optuna[28], a framework for hyperparameter optimization.

2.3.1. Convolutional Neural Network

This study introduces a custom CNN Classifier for sequence prediction, employing a multi-layer convolutional neural network architecture. The model consists of three convolutional layers, each with a varying number of filters (79, 111, and 180 filters, respectively) to extract spatial features from the input sequences. Max-pooling layers follow the first two convolutional layers to reduce the spatial dimensions of the feature maps, and batch normalization is applied after each convolutional layer to stabilize training. The flattened feature maps are then passed into fully connected layers consisting of 512 and 256 neurons to learn higher-level feature representations. Dropout is applied between the fully connected layers to prevent overfitting, and the final output layer produces multi-class predictions using a log softmax activation function. This model efficiently captures both local and global feature representations through the combination of convolutional and fully connected layers.

2.3.2. Recurrent Neural Networks

This study introduces a custom RNN Classifier for sequence prediction, utilizing a multi-layer RNN architecture. The model incorporates an RNN with multiple layers to capture temporal dependencies in the input sequences. The hidden layers of the RNN consist of a specified number of units, optimized for extracting sequential features. A batch normalization layer is applied to the output of the last time step to stabilize training and improve convergence. The model's sequential output is further passed through fully connected layers consisting of 512 and 256 neurons, respectively, to capture higher-level feature representations. Dropout is applied between the fully connected layers to reduce overfitting. This model efficiently captures both temporal patterns and higher-order representations through its recurrent and fully connected layers.

2.3.3. Long Short-Term Memory Network

The LSTM Classifier we used, incorporates LSTM layers that are designed to capture long-term dependencies in the input sequences by considering both past and future information at each time step. Specifically, the model first processes input sequences through a stacked LSTM layer containing 427 LSTM units, allowing it to retain relevant information across time steps. The LSTM is set with multiple layers (num_layers = 5) to improve its capacity to capture complex temporal relationships. The output from the final LSTM layer, representing the last time step, is passed through a dropout layer to mitigate overfitting. This is followed by a fully connected layer, which transforms the LSTM output into multi-class predictions. The model effectively leverages the sequential nature of LSTMs.

2.3.4. Bidirectional LSTM

The Bi-LSTM Classifier extends the capabilities of traditional LSTMs by incorporating bidirectional LSTM layers. These layers are uniquely designed to process information both forwards and backwards along the input sequence, enabling the model to capture context from both the past and future at each time step. This dual-direction processing allows for a more comprehensive understanding of temporal relationships within the data. The Bi-LSTM architecture features a bidirectional LSTM layer with 250 hidden units distributed across 4 layers, ensuring that each time

step is enriched with full sequence context. Such an arrangement not only enhances the model's ability to discern patterns that a single-direction LSTM might miss but also improves predictive accuracy by integrating insights from the entire sequence breadth.

2.3.5. Combined architectures

CNNs are particularly adept at capturing spatial information, making them highly effective for tasks that involve the analysis of grid-like data structures, such as images or sequences with local spatial dependencies. Additionally, LSTMs excel in handling sequential data, allowing for the modeling of long-term dependencies and temporal patterns, which are crucial in many predictive tasks. The integration of these two models aims to create a model that synergizes the spatial feature extraction capabilities of CNNs with the sequence modeling strengths of LSTMs. To explore the potential of this combination, we investigated several architectural strategies:

a) **CNN followed by LSTM (CNN-LSTM)**:
This architecture begins with the CNN, which extracts spatial features from the raw protein sequences. The CNN is adept at identifying local patterns or motifs within the sequence data that may be associated with cell aging. These spatial features are then fed into the Bidirectional LSTM, which captures the sequential dependencies among the CNN-extracted features. The Bidirectional LSTM's ability to analyze the sequence data from both directions ensures that critical temporal information is not overlooked.

b) **Parallel LSTM and CNN (pLSTM-CNN)**:
In this approach, the Bidirectional LSTM and CNN operate simultaneously on the input data. The Bidirectional LSTM processes the sequence data to capture temporal dependencies from both past and future directions, while the CNN extracts spatial features independently. The outputs of both models are then combined to form a comprehensive representation of the protein sequences, which is used for the final classification. This parallel architecture is designed to maximize the strengths of both models, capturing a wide range of information from the data by considering both spatial and sequential aspects simultaneously.

2.4. Evaluation metrics

The following commonly used evaluation metrics were employed to assess the effectiveness of the models[29]: sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC)[30], accuracy (Acc), F1 score, and AUC. The formulas for these metrics are provided below:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ score = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Where:

1) True Positive (TP): The number of correctly identified cases of cell aging.
2) True Negative (TN): The number of correctly identified non-aging cases.
3) False Positive (FP): The number of non-aging cases incorrectly identified as aging.
4) False Negative (FN): The number of aging cases incorrectly identified as non-aging.

AUC represents the area under the receiver operating characteristic (ROC) curve and is used to evaluate the overall performance of the model across different thresholds.

3. Results and discussion

3.1. ESM model selection

As detailed in the Methods section, we evaluated several variants of the Evolutionarily Scaled Model (ESM) for feature extraction from our protein sequence dataset. The goal was to identify the model that would provide the most accurate features while maintaining computational efficiency. Here, we present the results of this evaluation and the rationale for selecting the optimal ESM variant. The ESM variants evaluated included ESM-2_t6, ESM-2_t12, ESM-2_t30, and ESM-2_t33, each varying in complexity and embedding dimensions. The performance of these models was assessed based on their average accuracy across both the CNN and LSTM models, as well as the time required for feature extraction（Table 1）.

Table 1: Comparison of ESM-2 Model Variants for Feature Extraction

Model	#Layers	#Params(M)	#Emb Dim	Acc(%)	Time
ESM-2_t6	6	8	320	79.50	1
ESM-2_t12	12	35	480	**81.37**	2
ESM-2_t30	30	150	640	78.71	4
ESM-2_t33	33	650	1280	79.93	7

Note: 1. Time: The "Time" column represents the relative feature extraction time for each ESM-2 model variant. The extraction time for ESM-2_t6 is set as the baseline (1), and the times for other models are scaled relative to it. For example, ESM-2_t12 takes twice as long as ESM-2_t6, ESM-2_t30 takes four times as long, and ESM-2_t33 takes seven times as long.

The ESM-2_t12 model achieved the highest accuracy at 81.37%, outperforming the other variants. This accuracy advantage, coupled with its moderate computational demands, makes ESM-2_t12 the most effective choice for feature extraction. Therefore, we have selected ESM-2_t12 for use throughout our experiments. To further support the selection of the ESM-2_t12 model, we conducted a t-SNE analysis to visualize the separability of the features extracted by various methods. The t-SNE plot for ESM-2_t12 shows distinct clustering of data points based on their original class labels (Fig. 3A), indicating that ESM-2_t12 effectively captured meaningful features for classification. In contrast, the t-SNE plots for the traditional manual feature extraction methods,

including ACC(Fig. 3B), CKSAAP(Fig. 3C), and DDE(Fig. 3D), demonstrate less defined and more overlapping clusters. For ACC (Fig. 3B), while some degree of separation is visible, the clusters are not as distinct as those from ESM-2_t12, suggesting that ACC captures less discriminative features. CKSAAP (Fig. 3C) shows even more overlap between clusters, indicating a lower ability to distinguish between different classes. Finally, DDE (Fig. 3D) exhibits the least separability, with substantial overlap between data points from different classes, reflecting weaker feature extraction capabilities. The clear separation between clusters in the ESM-2_t12 model confirms the robustness of the extracted embeddings, reinforcing our decision to use this model for subsequent experiments.

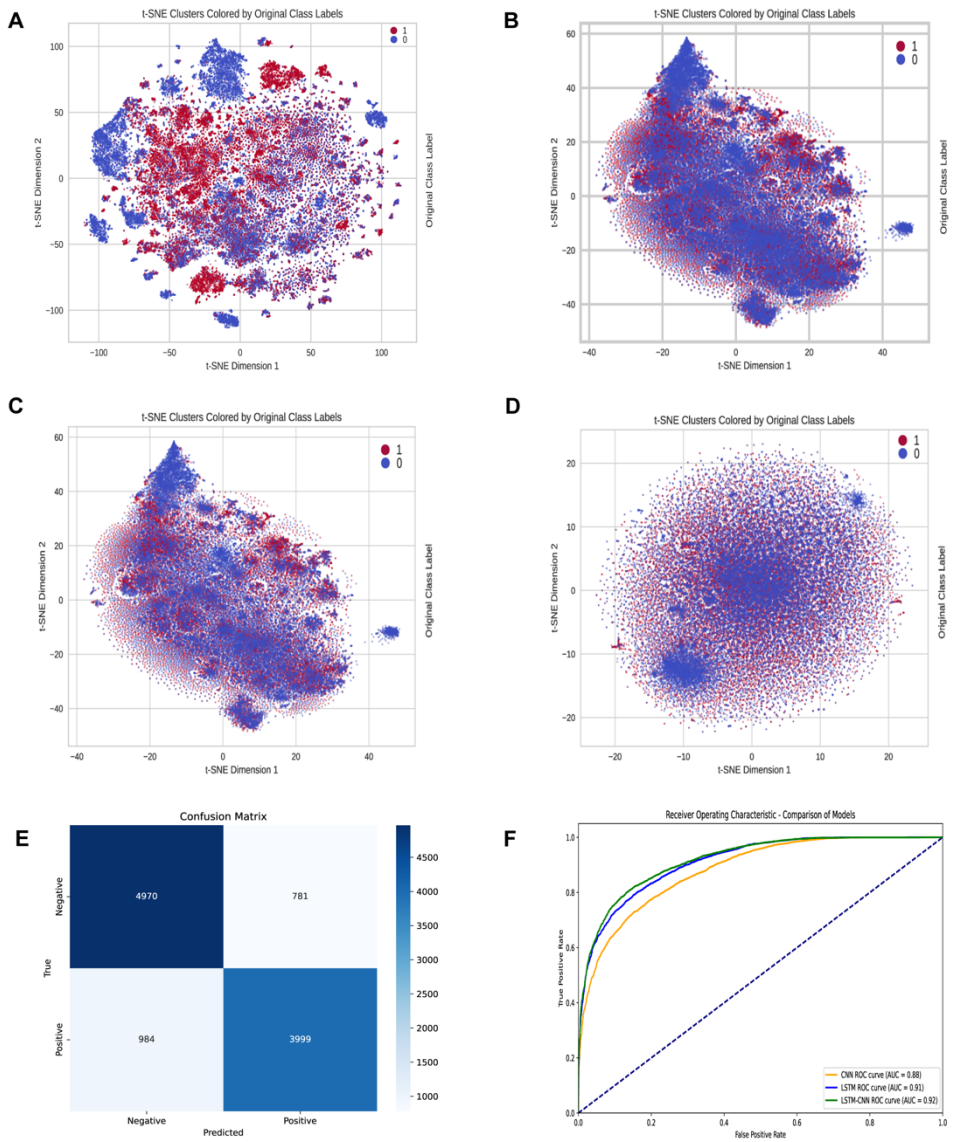

Fig 3. t-SNE visualization for ESM-2 and traditional methods A-D: A. t-SNE visualization for the ESM-2 model. B. t-SNE visualizetion for the ACC. C. t-SNE visualizetion for the CKSAAP. D. t-SNE visualizetion for the DDE. E. Confusion matrix for the SenSeqNet, illustrating a balanced classification outcome. F. Receiver Operating Characteristic (ROC) curves for the CNN, LSTM,

and SenSeqNet models, comparing true positive rates across various thresholds. The curves highlight the superior sensitivity and specificity of SenSeqNet, which effectively combines the strengths of both CNN and LSTM architectures.

3.2. Comparative performance evaluation of SenSeqNet with other deep learning models

We compared the performance of SenSeqNet with several deep learning models, including CNN, RNN, LSTM, Bidirectional LSTMs and various hybrid models. Each model was tested to determine its effectiveness in detecting cell aging from protein sequences, focusing on key performance metrics such as accuracy, AUC, Sn, Sp, and MCC using an independent test est. The comparative results of these models are summarized in Table 2**.**

Table 2: The performance comparison of SenSqeNet with other deep learning models on the independent test sets

Model	Acc%	Sn%	Sp%	F1-Score%	MCC
CNN	80.65	77.52	83.36	78.81	0.6104
RNN	81.12	77.86	83.93	79.29	0.6173
LSTM	82.09	77.48	86.09	80.07	0.6397
BiLSTM	82.53	**81.04**	83.83	81.16	0.6488
CNN-LSTM	78.02	72.85	82.47	74.16	0.5570
pLSTM-CNN	80.94	78.30	83.22	79.23	0.6164
SenSeqNet	**83.55**	79.53	**87.03**	**81.92**	**0.6689**

The CNN, known for its ability to capture spatial features, achieved an accuracy of 80.65% and an MCC of 0.6104. While effective in identifying local patterns within the protein sequences, the CNN struggled to capture temporal dependencies critical for detecting cellular senescence. To address this limitation, RNNs were introduced to model sequential data. The RNN showed an improvement, with an accuracy of 81.12% and an MCC of 0.6173, indicating its capability to better capture sequential patterns in the data. Next, more advanced temporal models, such as LSTMs, were explored. The LSTM, designed to handle long-term dependencies in sequential data, further improved accuracy to 82.09%, with an MCC of 0.6397. Building on this, the Bidirectional LSTM, which processes information in both forward and backward directions, enhanced the ability to capture complex dependencies, achieving an accuracy of 82.53% and an MCC of 0.6488. This model excelled at processing temporal patterns but lacked focus on spatial feature extraction.We also evaluated several hybrid architectures. Among these, the CNN-LSTM configuration demonstrated the lowest performance, with an accuracy of 78.02% and an MCC of 0.5570, suggesting that processing spatial features before sequential features may not fully leverage the strengths of both architectures. The pLSTM-CNN model performed better, achieving an accuracy of 80.94% and an MCC of 0.6164, though it had higher computational demands due to the simultaneous processing of spatial and sequential features.

The best-performing architecture was the SenSeqNet, which effectively combined the LSTM's ability to capture sequence dependencies with the CNN's proficiency in refining spatial features. SenSeqNet achieved the highest accuracy of 83.55%, an AUC of 0.92, and an MCC of 0.6689.

These results demonstrate that the combination of LSTM and CNN architectures, when sequenced in this order, leads to a more comprehensive understanding of protein sequences, resulting in superior predictive accuracy. Overall, the results highlight that while single models such as CNNs and LSTMs are effective in their respective domains of spatial or temporal feature extraction, the hybrid approach of SenSeqNet provides the best of both worlds, outperforming both individual models and other hybrid configurations. The confusion matrix and ROC curve (Fig. 3E, F) further confirm the stability and generalizability of SenSeqNet in detecting cellular senescence.

3.3. Comparative analysis of SenSeqNet with machine learning models

To further evaluate the performance of SenSeqNet, we also compared its performance with various commonly used machine learning models including RF, XGBoost, SVM, and Bagging. The highest accuracy achieved among these machine models was 78.05% by RF, while XGBoost showed a notably high Sn of 87.67% but compromised Sp at 67.06%. Overall, none of the traditional models surpassed 80% in accuracy, with moderate MCC ranging between 0.5318 and 0.5583. The results are summaried in Table 3

Table 3: Comparative analysis of SenSeqNet with machine learning models on the independent test set.

Model	Acc%	Sn%	Sp%	F1-Score%	MCC
RF	78.05	77.11	79.37	80.39	0.5583
XGBoost	74.83	**87.67**	67.06	72.42	0.5318
SVM	77.02	79.01	74.79	78.38	0.5386
Bagging	77.98	77.51	78.60	80.14	0.5565
SenSeqNet	**83.55**	79.53	**87.03**	**81.92**	**0.6689**

In comparison, the SenSeqNet demonstrated superior performance, achieving accuracy of 83.55%, highlighting its effectiveness in extracting complex, non-linear features from large-scale protein sequence data. This disparity in performance underscores the capability of SenSeqNet to better handle the extensive and intricate relationships inherent in biological datasets, especially when dealing with large volumes of data, where traditional machine learning models may struggle to fully capture contextual dependencies.

4. Conclusion

In this study, we have demonstrated the efficacy of a novel hybrid deep learning model SenSeqNet, which combining LSTM with CNN for the detection of cellular senescence from protein sequences. Our approach successfully leverages the sequential modeling capabilities of LSTMs and the spatial feature extraction strengths of CNNs to capture both temporal and spatial patterns within protein sequences that are indicative of cellular senescence. The SenSeqNet proved to be the optimal model, achieving an impressive accuracy of 83.55% and an AUC of 0.92 on the independent test set. The SenSeqNet outperformed both single models and other hybrid configurations, underscoring the importance of sequence processing order in deep learning architectures. To further test SenSeqNet's reliability, we conducted 10-fold cross-validation with a final cross-validated accuracy of 83.28% ± 0.58% that confirmed the SenSeqNet's model's reliability. Our findings highlight the potential of

deep learning in advancing the study of aging by enabling the accurate detection of aging at the cellular level. The robustness and scalability of the SenSeqNet model make it a valuable tool for future research, offering insights that advance the use of deep learning in aging studies while paving the way for identifying aging biomarkers and developing targeted therapeutic interventions. By leveraging advanced deep learning techniques and visualization tools, we aim to enhance our understanding of the molecular mechanisms of aging, ultimately contributing to strategies that promote healthy aging.

However, a key limitation of this study is the model's interpretability, particularly in relation to the identification of biologically significant motifs within the protein sequences. Unlike more extensively studied proteins, such as antibodies, where there is ample experimental validation and established motifs, the proteins associated with cell aging lack such data. This absence makes it difficult to verify whether the model is focusing on critical patterns that are biologically meaningful. The lack of known motifs in the context of cell aging constrains our ability to assert that the model's predictions are based on genuine biological markers, limiting the interpretability and potential clinical applicability of the results. Future research should aim to bridge this gap by exploring and validating potential motifs related to cell aging, thereby enhancing the model's utility and providing deeper insights into the biological mechanisms of aging.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no potential conflicts of interest.

Fundings

The work was supported by General Program of the National Natural Science Foundation of China (No. 82171622).

Authors' contributions

Hanli Jiang and Li Lin were responsible for the conceptualization and methodology development of the study, as well as drafting the original manuscript. Dongliang Deng and Siyi Liu contributed to data curation and investigation. Jianyu Ren and Xin Yang were involved in model development and visualization. Lubin Liu provided supervision, acquired funding, and led the review and editing of the manuscript. All authors have reviewed and approved the final version of the manuscript.

Acknowledgements

We extend our sincere appreciation to the Program for Youth Innovation in Future Medicine at Chongqing Medical University for their valuable support.

Availability of data and material

Publicly available datasets were analyzed in this study. The code utilized in this study is available at: https://github.com/HanliJiang13/SenSeqNet/.

Reference

1. Flatt T. A new definition of aging? Front Genet. 2012;3:148. doi: 10.3389/fgene.2012.00148
2. Childs BG, Durik M, Baker DJ, van Deursen JM. Cellular senescence in aging and age-related disease: from mechanisms to therapy. Nat Med. 2015;21:1424-1435. doi: 10.1038/nm.4000
3. Chen Z, Ain NU, Zhao Q, Zhang X. From tradition to innovation: conventional and deep learning frameworks in genome annotation. Brief Bioinform. 2024;25. doi: 10.1093/bib/bbae138
4. Lin L, Long Y, Liu J, Deng D, Yuan Y, Liu L, Tan B, Qi H. FRP-XGBoost: Identification of ferroptosis-related proteins based on multi-view features. Int J Biol Macromol. 2024;262:130180. doi: 10.1016/j.ijbiomac.2024.130180
5. Ashiqur Rahman S, Giacobbi P, Pyles L, Mullett C, Doretto G, Adjeroh DA. Deep learning for biological age estimation. Brief Bioinform. 2021;22:1767-1781. doi: 10.1093/bib/bbaa021
6. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. 2021;13:152. doi: 10.1186/s13073-021-00968-x
7. Zhou X, Zheng W, Li Y, Pearce R, Zhang C, Bell EW, Zhang G, Zhang Y. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. Nat Protoc. 2022;17:2326-2353. doi: 10.1038/s41596-022-00728-0
8. Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. Machine-learning methods for ligand-protein molecular docking. Drug Discov Today. 2022;27:151-164. doi: 10.1016/j.drudis.2021.09.007
9. Jiao S, Ye X, Sakurai T, Zou Q, Liu R. Integrated convolution and self-attention for improving peptide toxicity prediction. Bioinformatics. 2024;40. doi: 10.1093/bioinformatics/btae297
10. Dastider AG, Sadik F, Fattah SA. An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound. Comput Biol Med. 2021;132:104296. doi: 10.1016/j.compbiomed.2021.104296
11. Tsukiyama S, Hasan MM, Fujii S, Kurata H. LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec. Brief Bioinform. 2021;22. doi: 10.1093/bib/bbab228
12. Du Z, Ding X, Xu Y, Li Y. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. Brief Bioinform. 2023;24. doi: 10.1093/bib/bbad135

13. Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Brief Funct Genomics*. 2021;20:61-73. doi: 10.1093/bfgp/elaa030

14. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379:1123-1130. doi: 10.1126/science.ade2574

15. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*. 2008;9.

16. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49:D480-d489. doi: 10.1093/nar/gkaa1100

17. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26:680-682. doi: 10.1093/bioinformatics/btq003

18. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444. doi: 10.1038/nature14539

19. Nguyen VTD, Hy TS. Multimodal pretraining for unsupervised protein representation learning. *Biol Methods Protoc*. 2024;9:bpae043. doi: 10.1093/biomethods/bpae043

20. Mall R, Singh A, Patel CN, Guirimand G, Castiglione F. VISH-Pred: an ensemble of fine-tuned ESM models for protein toxicity prediction. *Brief Bioinform*. 2024;25. doi: 10.1093/bib/bbae270

21. Zhang F, Zhang Y, Zhu X, Chen X, Lu F, Zhang X. DeepSG2PPI: A Protein-Protein Interaction Prediction Method Based on Deep Learning. *IEEE/ACM Trans Comput Biol Bioinform*. 2023;20:2907-2919. doi: 10.1109/tcbb.2023.3268661

22. Müller AT, Hiss JA, Schneider G. Recurrent Neural Network Model for Constructive Peptide Design. *J Chem Inf Model*. 2018;58:472-479. doi: 10.1021/acs.jcim.7b00414

23. Khademi Z, Ebrahimi F, Kordy HM. A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals. *Comput Biol Med*. 2022;143:105288. doi: 10.1016/j.compbiomed.2022.105288

24. Breiman L. Bagging Predictors. *Machine Learning*. 1996;24:123-140. doi: 10.1023/A:1018054314350

25. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intelligent Systems and their Applications*. 1998;13:18-28. doi: 10.1109/5254.708428

26. Breiman L. Random Forests. *Machine Learning*. 2001;45:5-32. doi: 10.1023/A:1010933404324

27. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: Association for Computing Machinery; 2016:785 – 794.

28. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK, USA: Association for Computing Machinery; 2019:2623 – 2631.

29. Ao C, Jiao S, Wang Y, Yu L, Zou Q. Biological Sequence Classification: A Review on Data and General Methods. *Research (Wash D C)*. 2022;2022:0011. doi: 10.34133/research.0011

30. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21:6. doi: 10.1186/s12864-019-6413-7