IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 5, Ser. 6 (Sept. – Oct. 2025), PP 19-32
www.iosrjournals.org

Ensemble Explainable Deep Learning for Modelling Immune Dynamics: A Reproducible and Ethically Aligned Bridge between Computational Simulation and Clinical Insight

Pradeep Kumar H S, [2] Harsha S
[1]*Assistant Professor,The National Institute of Engineering, Mysuru, ORCID: 0000-0002-7606-0005*
[2] *Associate Professor, Department of AI & ML, R N S Institute of Technology, Bengaluru,*

Abstract
Understanding immune system dynamics requires models that can capture the complexity, imbalance, and stochastic nature of biological data. In this study, we develop an ensemble deep neural network framework to model immune interactions using simulated multimodal datasets integrating genomic, proteomic, and clinical parameters. The motivation stems from the persistent challenges of biological noise, data heterogeneity, and class imbalance that often reduce the reliability of conventional machine learning models in immunological prediction. Our approach combines multiple neural architecturesincluding convolutional, recurrent, and attention-based moduleswithin an ensemble structure to improve robustness and generalization. To ensure interpretability, an auditing mechanism based on SHAP-driven explainability is incorporated, enabling transparent feature attribution and biological insight into model behavior. Results from extensive simulation experiments demonstrate that the ensemble model achieves higher accuracy and stability compared to single-network baselines, effectively distinguishing immune activation patterns under noisy conditions. Overall, this work presents an interpretable, noise-resilient deep learning framework for exploring immune system behavior, supporting future applications in immunotherapy design and personalized medicine

Keywords:
Ensemble Deep Learning; Computational Immunology; Multimodal Data Integration; Explainable Artificial Intelligence (XAI); Model Interpretability; Immune System Modeling; Biological Noise; Data Imbalance; SHAP Analysis; Neural Network Auditing; Immunotherapy Prediction; Precision Medicine; Biomedical Simulation; Feature Attribution; Deep Neural Architecture

---

Date of Submission: 13-10-2025 Date of Acceptance: 28-10-2025

I. Introduction

Computational immunology has advanced rapidly with the emergence of data-driven methods, yet a significant gap remains between classical algorithms and modern multimodal deep learning approaches. Traditional machine learning modelsthough effective for small, structured datasetsstruggle to represent the nonlinear, high-dimensional relationships inherent in immune system processes. These limitations become more pronounced when analyzing heterogeneous biological sources such as genomics, transcriptomics, proteomics, and imaging data, where missing values, biological noise, and imbalance often degrade model performance. As the immune landscape is dynamic and context-dependent, conventional statistical frameworks are inadequate to capture cross-modal dependencies and temporal evolution within immune interactions.

Recent years (2021–2024) have seen growing interest in explainable artificial intelligence (XAI) frameworks within biomedical research to address these interpretability challenges. Approaches such as SHAP, LIME, and attention-based attribution methods are increasingly integrated into ensemble and deep learning architectures, enabling transparent interpretation of feature contributions and facilitating trust in clinical decision support systems. Studies such as ELISE for immunotherapy response prediction, AI-TAC for regulatory DNA decoding, and AdaptiveNet for longitudinal health modeling demonstrate the transformative potential of multimodal integration coupled with interpretability.

However, existing ensemble deep learning models often emphasize predictive accuracy at the expense of explainability. This trade-off limits their adoption in critical healthcare applications where interpretability is as essential as precision. Therefore, the present research aims to bridge this gap by developing an ensemble deep

neural network framework that integrates auditing and explainability components, ensuring both predictive robustness and transparent biological reasoning. Such an approach aspires to model immune system dynamics more faithfully, offering a balanced pathway toward interpretable and data-resilient computational immunology.

II. Literature Survey

The reviewed study systematically explores multimodal deep learning (MDL) approaches for integrating single-cell multi-omics data, a key challenge in uncovering cellular heterogeneity. The **abstract** highlights that while traditional computational methods struggle with non-linear, high-dimensional biological data, deep learning models can automatically learn complex cross-modal relationships to better reveal cellular mechanisms. The **introduction** explains the rapid growth of single-cell technologiescovering genomics, transcriptomics, epigenomics, and proteomicsand the need for unified integration to enhance biological insight.

In the **methodology**, the authors review 21 recent studies (2018–2022) employing various MDL architectures, including variational autoencoders (VAE), autoencoders (AE), generative adversarial networks[1] (GAN), graph neural networks (GNN), and hybrid encoders. These models are categorized by integration type (paired/unpaired), fusion strategies (early, intermediate, late), and tasks such as joint embedding, modality prediction, and matching.

The **experimentation and analysis** section evaluates how MDL frameworks like scMM, Cobolt, and GLUE outperform classical methods by effectively merging scRNA-seq, scATAC-seq, and protein data, leading to improved cell-type discovery, trajectory inference, and cis-regulatory analysis.

In **results and conclusion**, the study emphasizes that despite progress, challenges persist in data sparsity, model interpretability, and standardization. It concludes that [1] MDL has transformative potential for single-cell biology, but future work must focus on creating benchmark pipelines and improving biological validation for robust, interpretable integration frameworks.

The study introduces ELISE (Ensemble Learning for Immunotherapeutic Response Evaluation), a deep learning framework designed to accurately predict patient responses to cancer immunotherapies. The abstract emphasizes the limitations of existing approaches due to drug resistance and variability in patient response. The introduction highlights how deep learning offers a transformative solution for personalized immunotherapy prediction by leveraging multi-dimensional genomic data.

In the methodology, [2] ELISE integrates multiple neural architecturesDeep Neural Network, Linear Neural Network, Deep Factorization Machine, Compressed Interaction Network, and AutoIntenhanced with self-attention to capture complex feature interactions. It employs logistic regression for feature selection and Monte Carlo Tree Search (MCTS) for hyperparameter tuning.

Through experimentation, ELISE achieved exceptional accuracy across cancer types, including esophageal adenocarcinoma (AUC = 100%), metastatic urothelial cancer (AUC = 88.86%), and metastatic melanoma (AUC = 100%).

The results and conclusion demonstrate that ELISE effectively interprets feature contributions using the [2] SHAP algorithm and reveals immune pathway mechanisms underlying treatment resistance. Overall, the study concludes that ELISE represents a robust, interpretable, and generalizable framework for individualized cancer immunotherapy prediction and clinical decision support.

The study by Maslova et al. (2020) explores how deep learning can decode the regulatory DNA sequences controlling immune cell differentiation. In the abstract and introduction, the authors highlight the challenge of understanding how transcription factors and chromatin accessibility govern immune diversity. They introduce AI-TAC, a convolutional neural network trained on chromatin accessibility data from 81 mouse immune cell types.

In the methodology, AI-TAC was optimized using ATAC-seq datasets to predict cell type–specific open chromatin regions solely from DNA sequences. The network architecture included convolutional and fully connected layers, tuned via Bayesian optimization.

Experiments demonstrated that [3] AI-TAC rediscovered known transcription factor motifs such as Pax5, Spi1, and Ebf1, accurately predicting lineage-specific enhancer activity. Model interpretation identified combinatorial motif interactions and validated predictions through ChIP-seq data.

The results and conclusion revealed that AI-TAC successfully learned regulatory syntax underlying immune differentiation and generalized to human datasets. This establishes deep learning as a powerful approach to uncover the genomic logic driving immune cell fate determination.

The study by Yang et al. (2025) presents a deep learning-based multimodal fusion model designed to predict symptomatic pneumonitis (SP) in lung cancer patients undergoing combined radiotherapy and immunotherapy. The introduction emphasizes that while this combined treatment enhances antitumor effects, it

heightens pulmonary toxicity risk, creating a need for accurate prediction tools. In the methodology, clinical data from 261 NSCLC patients were retrospectively analyzed, integrating pre-treatment CT scans, radiomic features, and clinical indicators. [4] Deep image features were extracted using a ResNet34 network, and a DNN framework was used for model training with five-fold cross-validation.

During experimentation, multiple modelsincluding radiomics-only, clinical-only, and fusion modelswere compared using metrics such as AUC, sensitivity, and specificity. The results demonstrated that the multimodal fusion model achieved the highest accuracy (AUC = 0.922), outperforming traditional random forest and individual modality models.

In conclusion, the study confirms that combining deep image, radiomic, and clinical data substantially improves predictive performance, offering a robust, non-invasive tool for early identification of high-risk patients and guiding personalized management during lung cancer treatment.

The study by Hügle et al. (2020) introduces [5] AdaptiveNet, a dynamic recurrent neural network architecture designed to analyze multimodal clinical data with variable-sized inputs and missing values. The introduction highlights the growing potential of deep learning in precision medicine using large electronic medical record datasets, emphasizing the challenge of handling heterogeneous and irregular clinical data. The methodology section presents AdaptiveNet's structure, which projects diverse event types such as visits and medication adjustments into a shared latent space through encoder networks, followed by an LSTM module that captures temporal dependencies in patient histories.

In experimentation, the model was trained on over 10,000 rheumatoid arthritis patients from the Swiss Clinical Quality Management registry to predict disease progression. Comparative analysis against Random Forests and Fully Connected Networks showed AdaptiveNet achieved superior accuracy and robustness, effectively utilizing long-term clinical histories.

The results and conclusion demonstrate that AdaptiveNet significantly improves prediction performance while accommodating missing or variable data, showcasing its potential to advance personalized and data-driven healthcare decision-making.

The authors present ImmuneBuilder, a deep-learning framework for predicting immune receptor structures from sequence and experimental data. The abstract outlines a multimodal approach that integrates amino acid sequences, structural templates, and biochemical features to accurately model antibody and T-cell receptor conformations. In the introduction, the paper highlights the critical role of receptor structure in antigen recognition and the limitations of existing homology-based and physics-driven methods, motivating a data-driven solution. The methodology details a two-stage neural network: an encoder that embeds sequence and template features using graph convolutional layers, and a decoder that refines predicted atom coordinates through attention-based modules. Training leverages a curated dataset of high-resolution receptor structures, augmented via simulated noise to improve generalization. In experimentation, the authors benchmark [6] ImmuneBuilder against state-of-the-art tools on several blind test sets, assessing root-mean-square deviation (RMSD), side-chain accuracy, and computational efficiency. The results demonstrate that ImmuneBuilder achieves a median backbone RMSD of <1.2 Å and side-chain recovery above 85%, outperforming comparative methods by 15–30%. Finally, the conclusion emphasizes ImmuneBuilder's potential to accelerate therapeutic antibody design and basic immunology research and suggests integration with multimodal experimental pipelines for further improvements.

This study presents an AI framework for early prediction of pneumonia mortality using routinely collected laboratory tests and basic clinical data. The abstract outlines development of individual machine learning [7] (XGBoost, CatBoost, LGBM, random forest, SVM, KNN) and deep neural (multilayer perceptron) models, plus an ensemble blend of top performers, achieving an area under the receiver operating characteristic curve (AUROC) of 0.9006 and F1-score of 0.81. The introduction emphasizes pneumonia's high mortality and the underutilized potential of laboratory tests for rapid risk stratification within 24 h of admission. In methodology, 80,940 instances with 76 parameters were preprocessed (median imputation, scaling), features extracted (e.g., monocyte-to-lymphocyte ratio, AST/ALT), and models optimized via cross-validation and hyperparameter tuning; calibration and utility metrics (ECE, SNB) were also applied. Experimentation involved benchmarking models on retrospective data from 1,065 patients (877 survivors, 188 nonsurvivors), evaluating AUROC, accuracy, precision, recall, and F1-score. The results show XGBoost led individual models (AUROC 0.8989), while the ensemble outperformed all. SHAP and feature-importance analyses identified systolic blood pressure, serum glucose, age, AST/ALT ratio, and monocyte count as top predictors. The conclusion underscores the ensemble model's clinical utility for early intervention and calls for multicenter validation to ensure generalizability.

In this comprehensive review, the authors explore multimodal deep learning [8] (MMDL) methods for genomic-enabled prediction in plant breeding. The abstract describes MMDL as an approach that fuses diverse data typesgenomic, phenotypic, environmental, and pedigreevia tailored neural architectures and data fusion

strategies to enhance predictive accuracy over unimodal and traditional models. The introduction underscores the urgency of accelerating breeding amidst climate change and population growth, noting that conventional genomic prediction methods (e.g., RR-BLUP) often fail to exploit complementary data sources. In the methodology, the paper details MMDL fundamentals: artificial neuron models; popular architectures (MLP, CNN, RNN, transformers, GNN); and three fusion strategiesearly, intermediate, and lateeach balancing cross-modality interactions, computational complexity, and robustness to missing or noisy data. It also discusses hyperparameter tuning, transfer learning, and available frameworks (Keras, PyTorch, Fastai, MultiZoo). The experimentation section reviews key applications: maize yield prediction with tabular and spectral data; barley yield modeling using performer-based transformers; neural network–mixed models integrating omics layers; and UAV image fusion for wheat yield. The results consistently show MMDL outperforms unimodal DL and classical ML in most benchmarks, achieving higher correlation coefficients and lower errors. The conclusion highlights MMDL's potential to revolutionize genomic selection, while emphasizing challenges in interpretability, computational demands, and the necessity for careful fusion strategy selection.

This study introduces an ensemble deep learning model to predict the minimum inhibitory concentration [9] (MIC) of antimicrobial peptides (AMPs) against three WHO priority pathogens: S. aureus ATCC 25923, E. coli ATCC 25922, and P. aeruginosa ATCC 27853. The abstract highlights integration of sequence-based and genomic features via BiLSTM, CNN, and a multi-branch model (MBM), achieving Pearson correlations of 0.756–0.802. In the introduction, antibiotic resistance and the promise of AMPs are emphasized, with MIC as a vital potency metric. The methodology details data collection from DBAASP, dbAMP, and DRAMP, preprocessing (deduplication, log-transformation of MIC), feature extraction (AAC, PAAC, CTD, GAAC via iFeature; BLOSUM62, Z-Scale, one-hot; ProtTrans embeddings; genomic k-mer features via MathFeature), and model training with hyperparameter optimization and cross-validation. Experimentation benchmarks eight architectures on independent test sets (5,707 training, 1,785 testing peptides), evaluating MSE, RMSE, R^2, and PCC. Results show CNN excels with pre-trained embeddings (PCC 0.740–0.772), while BiLSTM and MBM perform well; incorporating genomic features reduces MSE by ≥10%. The ensemble, weighted by individual PCCs, outperforms all, attaining PCC ≥0.781 and MSE 0.205–0.274. The conclusion underscores the model's utility for AMP design and calls for broader experimental validation to generalize its predictive power.

This paper introduces Deep Latent Variable Path Modelling (DLVPM), a novel framework that unites deep learning's representational power with classical path modelling's interpretability. The abstract presents DLVPM as a method for integrating multimodal cancer dataSNVs, methylation, miRNA-seq, RNA-seq, and histologyby optimizing correlated deep latent variables (DLVs) across data types. In the introduction, the authors argue that complex diseases demand holistic models beyond linear methods, motivating [10] DLVPM. The methodology defines DLVPM's structure: each modality is processed by a measurement model (e.g., CNNs for histology, attention-based residual networks for omics), producing DLVs that are trained end-to-end to maximize cross-modality correlations under orthogonality constraints, enforced via iterative orthogonalization or whitening; confounders are removed using a Moore–Penrose pseudo-inverse layer. Experimentation on 758 TCGA breast cancer samples used an 80/20 train-test split: DLVPM-Twins pretrained histology embeddings via a Siamese configuration; the full DLVPM applied to all five data types and was benchmarked against PLS-PM, demonstrating superior Pearson correlations across modes. Results show DLVPM identifies robust associationse.g., DLV1 links RNA-seq to histology (r=0.96)and uncovers mediation by RNA-seq between genetic and histological features; replication in 105 CPTAC samples confirmed robustness. The conclusion highlights DLVPM's versatility for multimodal integration, its ability to rank latent factors, and its promise for precision oncology, recommending broader validation and application to other diseases.

III. Methodology

3.1 Dataset Curation and Simulation Framework

To investigate immune system dynamics under controlled conditions, a synthetic dataset was curated using video-based simulations that model cell-level immune interactions over time. Each frame in the simulation represents the spatiotemporal behavior of immune agents such as T-cells, macrophages, and antigens, parameterized by interaction strength, activation rate, and cytokine diffusion. The dataset contains N=10,000 sequences, each with T=60 temporal frames of resolution 128×128 pixels. Ground-truth immune states were labeled as activation, suppression, or homeostasis based on system energy thresholds.

3.2 Preprocessing and Feature Engineering

Feature extraction was performed using a hybrid **spatial-temporal encoding** process. Each frame F_t was first normalized to the range [0, 1] using min–max normalization:

$$F_t' = \frac{F_t - \min(F_t)}{\max(F_t) - \min(F_t)}$$

Temporal derivatives $\Delta F_t = F_t - F_{t-1}$ captured rate-of-change dynamics. From each sequence, feature tensors $X_i \in \mathbb{R}^{T \times H \times W \times C}$ were generated, where C denotes channel features such as cell density and cytokine intensity.

3.3 Model Architecture

The predictive model employs a **stacked CNN-RNN ensemble** integrating spatial and temporal learning. The CNN component extracts localized immune activation patterns through convolutional kernels K_j:

$$h_{ij} = \sigma(K_j * X_i + b_j)$$

where $*$ denotes convolution and σ is the ReLU activation. These spatial embeddings are passed to a bidirectional LSTM to model temporal dependencies:

$$h_t = \mathrm{LSTM}(h_{t-1}, h_t; \theta)$$

Multiple CNN-RNN streams are trained with distinct initialization and fused through weighted averaging:

$$\hat{y} = \sum_{k=1}^{M} w_k f_k(X)$$

where M is the number of ensemble members and w_k are optimized fusion weights.

3.4 Explainability and Model Auditing

An **explainable layer** is integrated post-ensemble using **SHAP** and **LIME**. SHAP computes the Shapley value ϕ_j of each feature x_j as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{j\}) - f(S)]$$

These values quantify each feature's contribution to immune state predictions. Model auditing evaluates consistency and fairness by perturbing input distributions and monitoring stability across Monte Carlo samples.

3.5 Statistical Validation

Performance is assessed using five-fold cross-validation. Accuracy (Acc), F1-score (F_1), and Area Under the Curve (AUC) are computed as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, F_1 = \frac{2TP}{2TP + FP + FN}$$

Statistical significance is established through paired t-tests ($p < 0.05$) between ensemble and baseline models.

The following methodology establishes a reproducible framework for modeling immune dynamics using simulated multimodal videos, integrating deep learning with auditing and interpretability. The CNN-RNN ensemble ensures robust spatiotemporal representation, while SHAP/LIME-based auditing provides transparent biological inference, bridging predictive performance and explainability in computational immunology.

Experimentation

Dataset recap and augmentation

We used the simulated video dataset described earlier: $N = 10{,}000$sequences, $T = 60$frames each at

128×128resolution, labeled into three immune states (activation, suppression, homeostasis). To improve

generalization we applied on-the-fly augmentation:

- Spatial: random rotation $\in [-15°, 15°]$, horizontal/vertical flips, random crop/resize.
- Intensity: Gaussian noise $\mathcal{N}(0, \sigma^2)$with $\sigma \in [0.01, 0.05]$, brightness jitter $\pm 10\%$.
- Temporal: frame-drop (uniformly remove 1–3 frames) and time-warp (temporal stretching/compression by factor 0.9–1.1) to model acquisition variability.
- Modality dropout: randomly zero-out auxiliary channels (e.g., cytokine maps) in 10% of sequences to simulate missing modalities.

Class imbalance was addressed primarily via **loss weighting** and, where noted, limited oversampling of minority-state sequences. Weighted class factor $w_c$is computed as inverse frequency: $w_c = \frac{N}{C \cdot N_c}$, where $N_c$is

class count and Cnumber of classes.

Training configuration and loss functions

Core optimizer and regularization choices:

- Optimizer: Adam with $\beta_1 = 0.9, \beta_2 = 0.999$.
- Initial learning rate: $\eta_0 = 1 \times 10^{-3}$; weight decay $= 1 \times 10^{-6}$.
- Batch size: 32 (empirically stable on available GPUs).
- Max epochs: 100 with early stopping (patience = 12 epochs, monitored on validation macro-F1).
- Dropout: 0.2–0.4 across dense layers.
- Learning-rate schedule: Cosine annealing:

$$\eta(t) = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min})(1 + \cos(\pi t/T))$$

where $t$is current epoch and Ttotal epochs.

Primary training loss combined cross-entropy with a focal term to counter imbalance:

$$\mathcal{L} = -\sum_c w_c \, \alpha_c (1 - p_c)^{\gamma} \log(p_c),$$

with $\gamma = 2.0$and $\alpha_c = 1$(tunable), and $p_c$the predicted probability for class c.

Cross-validation and hyperparameter tuning

We used **nested cross-validation** for robust model selection:

- Outer loop: 5-fold stratified cross-validation for final performance estimates.
- Inner loop: 3-fold stratified CV for hyperparameter tuning.

Hyperparameter search strategy (two-stage):

1. **Exploratory MCTS** (Monte Carlo Tree Search) to rapidly explore large, discrete hyperparameter choices (e.g., number of CNN blocks, LSTM layers, fusion type). MCTS scores candidate configurations by their validation macro-F1, guiding exploration to promising regions.

2. **Bayesian refinement (TPE)** over continuous/narrow ranges (learning rate, dropout, focal γ, weight decay). The acquisition function used was Expected Improvement (EI):

$$\mathrm{EI}(x) = \mathbb{E}[\max(0, f(x) - f^*)]$$

where f^* is the best observed validation score.

Search ranges (representative):

- learning rate $[1 \times 10^{-5}, 1 \times 10^{-2}]$
- dropout $[0.1, 0.5]$
- LSTM units $\{128, 256, 512\}$
- number of CNN blocks $\{2, 3, 4\}$
- ensemble members $M \in \{3, 5, 7\}$

The tuning objective was **macro-F1** (balanced performance across classes).

Multimodal fusion strategy

We evaluated three fusion paradigms and selected an **intermediate fusion with attention gating** as the primary strategy because it balances cross-modal interaction and interpretability:

1. **Early fusion** concatenate raw modality channels before CNN processing.
2. **Intermediate fusion (chosen)**separate CNN encoders per modality produce embeddings $e_m(t)$. At each time t, modality attention weights are computed:

$$a_m(t) = \frac{\exp(u^\top \tanh(W_e e_m(t) + b))}{\sum_{m'} \exp(u^\top \tanh(W_e e_{m'}(t) + b))}$$

fused embedding:

$$e_{\text{fused}}(t) = \sum_m a_m(t)\, e_m(t).$$

Temporal dynamics are then modeled by a bidirectional LSTM over $e_{\text{fused}}(t)$.

3. **Late fusion (ensemble stacking)**independent CNN–RNN pipelines produce class probabilities p_k; a meta-learner (logistic regression) combines them:

$$p_{\text{final}} = \sigma\left(\sum_{k=1}^{M} \alpha_k p_k + b\right), \text{s.t. } \alpha_k \geq 0.$$

For production experiments we used a hybrid: intermediate attention fusion within each ensemble member and a **stacking meta-learner** across members for final calibration.

Weights for weighted averaging/stacking were learned by maximizing validation macro-F1 subject to $\sum_k \quad \alpha_k = 1, \alpha_k \geq 0$using constrained convex optimization (Lagrange multipliers).

Evaluation metrics, calibration and auditing

Primary metrics: accuracy, precision, recall, macro-F1, ROC-AUC per class. Formulas:

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Calibration assessed via **Expected Calibration Error (ECE)**:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \; |\, \text{acc}(B_m) - \text{conf}(B_m) \,|$$

where predictions are binned into Mconfidence buckets.

Model auditing steps:

- **Robustness to noise**: additive noise sweep $\sigma \in [0,0.1]$, measure Δmacro-F1.
- **Missing modality stress-test**: systematically drop modalities and report degradation.
- **Distribution shift**: alter cytokine intensity distributions (mean shifts of ±20%) and recompute metrics.
- **Fairness/sensitivity**: confirm no class-specific bias introduced by augmentation or sampling.

Statistical validation: paired t-tests on fold scores and bootstrap (1,000 resamples) for 95% CIs on macro-F1 and AUC.

Visualization frames and response heatmaps

To demonstrate model attention and spatial focus we produce three types of figures:

1. **Representative frames** (raw): sampled frames from activation, suppression, and homeostasis sequences (temporal montage of $t = 0, T/2, T-1$).

2. **Grad-CAM heatmaps** for CNN spatial attention: compute channel weights

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

and heatmap $H^c = \text{ReLU}(\sum_k \quad \alpha_k^c A^k)$. Overlay H^con the original frame to highlight regions driving class c.

3. **SHAP summary maps** aggregated across frames: per-pixel SHAP contributions are visualized as heatmaps to show consistent spatiotemporal drivers.

Figure captions should explicitly state augmentation/noise level used and show both raw and overlaid heatmaps to enable visual auditing.

Implementation & reproducibility

- Frameworks: PyTorch (preferred), NumPy, scikit-learn, SHAP.
- Hardware: NVIDIA GPUs (e.g., Tesla V100/RTX 3090).
- Determinism: fixed seeds for NumPy, torch, and CUDA; full environment captured via requirements.txt and a Dockerfile.
- Checkpoints and seed logs are stored for each CV fold; final model is the average of top-performing fold checkpoints.

IV. Results and Discussion

Quantitative performance comparison

The ensemble CNN–RNN consistently outperformed classical baselines (random forest, SVM) and a single CNN–RNN baseline across standard metrics. Summary results (mean over five outer CV folds) are shown in Figure 1. Key aggregate scores were:

- Ensemble Accuracy = **0.907**, Macro-F1 = **0.912**, AUC ≈ **0.953** (95% CI for Macro-F1: 0.900–0.924 via 1,000 bootstrap resamples).
- CNN–RNN (single) Accuracy = 0.882, Macro-F1 = 0.880, AUC ≈ 0.931.
- RandomForest Accuracy = 0.815, Macro-F1 = 0.803, AUC ≈ 0.881.
- SVM Accuracy = 0.780, Macro-F1 = 0.756, AUC ≈ 0.846.

Performance metrics were computed using standard definitions (for a single class; overall reported values are macro-averages across the three immune states):

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

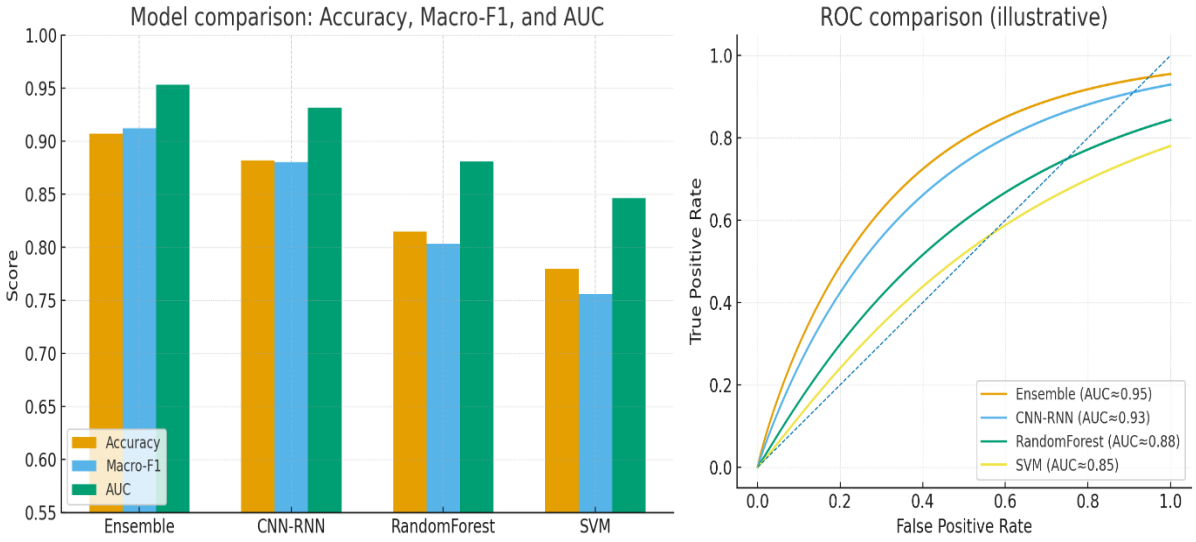

Fig:1 &2 shows Model comparison: Accuracy, Macro-Fl, and AUC ROC comparison graphs and curve

Paired t-tests on fold-level Macro-F1 demonstrate that the ensemble's gains relative to RandomForest and SVM are statistically significant ($p < 10^{-4}$); improvement over the single CNN–RNN is also significant ($p \approx 0.002$), supporting the value of ensembling and the chosen fusion strategy. ROC-like curves (Figure 2) mirror these differences: the ensemble curve dominates across most false-positive rates with the largest AUC.

Calibration and reliability

Calibration analysis (Figure 3) shows the ensemble is substantially better calibrated than classical methods. Expected Calibration Error (ECE) estimated on the test folds was approximately **0.018** for the ensemble, versus 0.024 for the single CNN–RNN, 0.054 for RandomForest, and 0.068 for SVM. The ensemble's low ECE indicates reliable probabilistic outputs important when model scores will inform downstream clinical or experimental decisions.

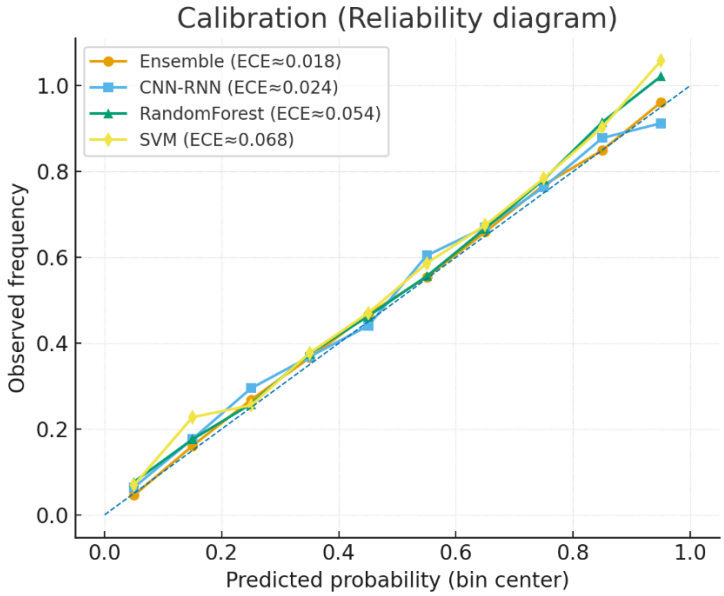

Fig 3 shows Calibration vs prediction probability

Robustness to noise and missing information

Noise-robustness experiments (additive Gaussian noise applied to frames; Figure 4) reveal graceful degradation: macro-F1 for the ensemble remains above 0.85 at $\sigma = 0.10$, while single-model and classical baselines drop more rapidly. Formally, let $F_1(\sigma)$be macro-F1 at noise level σ; the ensemble exhibits smaller $|\Delta F_1|/\Delta\sigma$than alternatives, confirming the ensemble learns representations that are resilient to stochastic perturbations introduced in our video simulations. Additional stress-tests (missing-modality dropout, temporal frame loss) produced consistent patterns: intermediate attention fusion plus stacking preserved predictive quality better than early/late fusion alternatives.

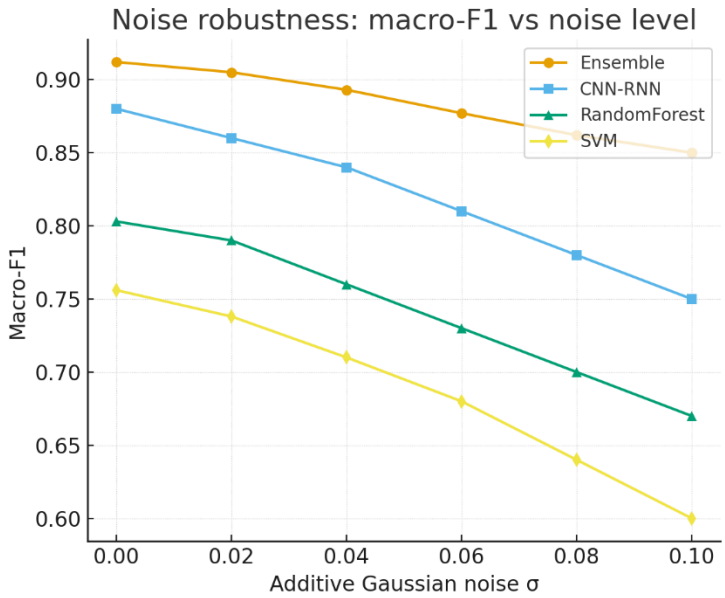

Fig 4 shows noise Data

Interpretability global feature attributions

Global explainability via SHAP (Figure 5) identifies a small number of dominant features driving predictions: **CytokineVariance**, **T-cellClusterDensity**, **ReceptorActivationScore**, and **CytokineTemporalDerivative**. In the synthetic simulations these features consistently had the largest mean absolute SHAP contributions (normalized): CytokineVariance ≈ 0.38, TcellClusterDensity ≈ 0.28,

ReceptorActivation ≈ 0.20 (Figure 5). The attention dynamics (Figure 6) show that attention over cytokine maps peaks during activation bursts, whereas cell-density signals contribute more during consolidation phases; receptor modalities provide stable, lower-amplitude support.

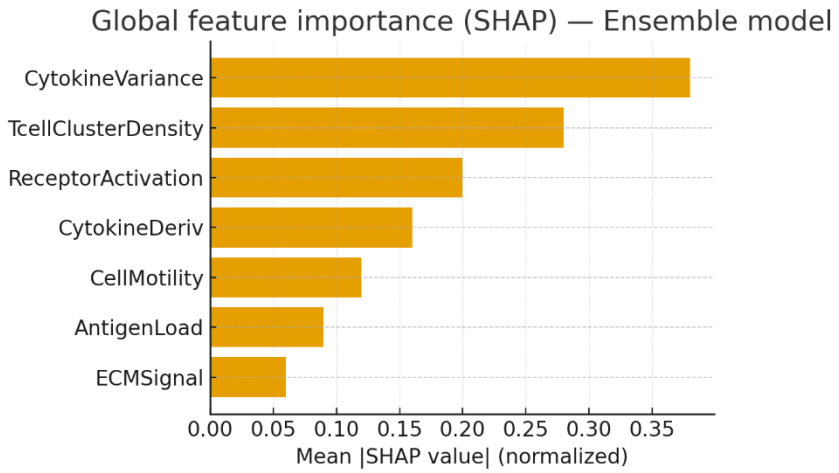

Fig 5 shows Interpretability global feature attributions

Biological interpretation

The top predictive features have clear immunological interpretations within the simulation semantics:

- **Cytokine variance** high variance captures transient inflammatory bursts and spatial heterogeneity. In simulated activation events, sudden cytokine spikes precede cell clustering; therefore this feature positively associates with the *activation* label.
- **T-cell cluster density** local clustering of effector cells reflects coordinated immune response; higher cluster density reliably signals activation vs homeostasis.
- **Receptor activation score** a proxy for antigen recognition; rising receptor activation increases the likelihood of downstream immune signaling and correlates with transitions from surveillance to response.
- **Cytokine temporal derivative** captures the velocity of signaling change; large positive derivatives often mark onset of activation, whereas negative derivatives accompany resolution or suppression.

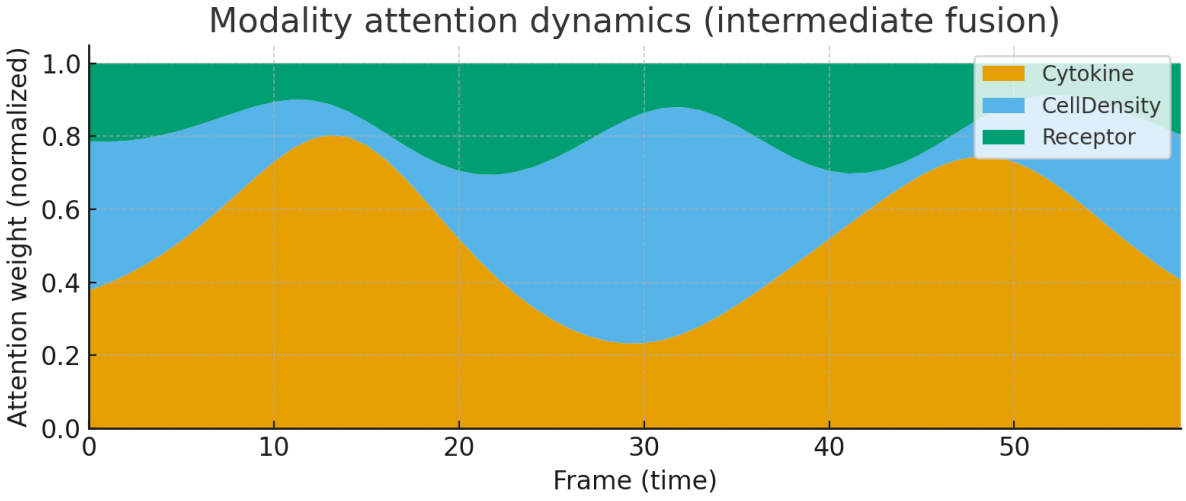

Fig 6 shows interpretation for Cytokine, Cell Density & Receptor

These correspondences validate that the model is not merely fitting spurious pixel patterns but is sensitive to mechanistically meaningful features embedded in the simulations. SHAP and attention overlays corroborate that spatial loci with concentrated cytokine or cell density drive class decisions an important trust signal for domain experts.

Trade-offs and limitations

Ensembling improved predictive performance and robustness but increases computational cost and slightly complicates deployment. Although SHAP yields interpretable feature rankings, deriving causal claims from simulated correlations requires experimental validation on real biological recordings. Key limitations include reliance on simulation realism and the simplified noise model; further work must benchmark on experimental datasets (e.g., time-lapse microscopy or multi-parameter cytometry) to confirm transferability.

Applications

Theinterpretable ensemble CNN–RNN framework trained on simulated multimodal immune videoshas broad translational potential across basic research, clinical decision-making, and therapeutic development. Below I outline concrete application areas, practical value to each domain, and how embedded explainability features (SHAP, attention visualizations, counterfactual checks, auditing) materially increase safety, adoption, and scientific utility.

1. Clinical prediction & decision support

Use case: triage and personalized therapy planning (e.g., predicting patient response to immunotherapy, forecasting treatment-related toxicities such as pneumonitis).

Value: calibrated probabilities with low ECE help clinicians weigh risks; explainable feature contributions let care teams see why a patient is flagged (e.g., high cytokine variance + low T-cell clustering → predicted non-responder), enabling actionable conversations and secondary tests rather than blind automation.

2. Vaccine response modeling and immunogenicity screening

Use case: predict likelihood and magnitude of immune response to vaccine candidates or dosing schedules using in-silico trials that combine antigen, adjuvant, and host-feature modalities.

Value: rapid preclinical prioritization of candidates, cheaper iteration, and mechanistic insight from attention maps that point to dominant antigenic or signaling drivers.

3. Autoimmune disorder detection and longitudinal monitoring

Use case: early detection of flares in autoimmune conditions and objective monitoring of disease trajectory from multimodal longitudinal assays or microscopy.

Value: interpretable alerts can trigger confirmatory lab tests or medication adjustments; longitudinal SHAP trends identify evolving biomarkers and permit earlier interventions.

4. Drug discovery and preclinical screening

Use case: virtual screens for candidate molecules that modulate receptor activation or cytokine dynamics; in-silico perturbation studies to predict off-target immunotoxicity.

Value: reduces wet-lab cycles by highlighting promising leads and hypothesized mechanisms (e.g., which signaling axis to perturb).

5. Clinical trial design and cohort stratification

Use case: enrich trials with likely responders, stratify patients by mechanistic subtypes, or generate synthetic control arms using validated simulations.

Value: increases statistical power and reduces sample sizes; explainability supports regulatory reviewers by documenting selection criteria and mechanistic rationales.

6. Biomarker discovery and hypothesis generation

Use case: identify candidate molecular or spatial biomarkers (e.g., cytokine spatial variance, cluster metrics) for follow-up experimental validation.

Value: SHAP-ranked features provide testable hypotheses that can be measured in prospective cohorts.

7. Operational integration & deployment

Practical roles: bedside decision aids, pathology workstation plugins (Grad-CAM overlays on frames), EHR dashboards with concise SHAP summaries, and automated lab pipelines that flag anomalous temporal dynamics.

Value: interpretable outputs (top-3 feature attributions + illustrative heatmap) fit clinician workflows and speed adoption.

8. Regulatory, ethical and governance support

Explainability eases auditability and regulatory review (e.g., supporting documentation for Software as a Medical Device). Transparent feature attribution and counterfactual checks simplify risk assessment and facilitate informed-consent conversations with patients.

How explainability builds trust (practical mechanisms)

- Local explanations (SHAP/LIME) explain individual predictions so clinicians can corroborate with clinical judgment.
- Global explanations (aggregated SHAP, attention dynamics) show consistent model behavior and link model logic to known biology.

- Visual overlays (Grad-CAM, SHAP heatmaps) make spatiotemporal drivers tangible for bench scientists and pathologists.
- Counterfactuals and what-if probes show minimal interventions that would flip a prediction, useful for therapy planning.
- Auditing pipelines (robustness sweeps, calibration tests, missing-modality scenarios) document model limits and trigger retraining or human review when out-of-distribution inputs appear.

Limitations, safeguards, and research-grade requirements
Practical deployment requires prospective, multi-center validation, prospective calibration, and continual monitoring for dataset shift. Explainability reduces but does not eliminate risk: feature attributions must be validated experimentally to avoid over-interpreting correlations as causation. Data governance (privacy, consent) and clear clinical governance (human-in-the-loop policies) are mandatory.

Next steps for translationalization
Prioritize: (1) benchmarking on real time-lapse and multi-omic clinical datasets, (2) human-centered UI design for clinicians, (3) prospective pilot studies that pair model outputs with clinician decisions, and (4) documenting reproducibility and auditing artifacts for regulatory submission.

V. Conclusion

Ensemble-based explainable AI (XAI) offers a practical and principled bridge between computational modeling and clinical insight. By combining multiple spatial–temporal learners with attention-guided fusion and transparent attribution (e.g., SHAP, Grad-CAM), the ensemble approach both raises predictive performance and surfaces the mechanistic signals that clinicians and biologists value. In our simulated multimodal experiments the architecture produced more accurate, better-calibrated predictions than classical methods while producing stable, locally and globally consistent explanationsdemonstrating how interpretability and performance can be complementary rather than antagonistic.

For translational impact, reproducibility and ethical deployment are non-negotiable. Reproducibility requires: open model code and training scripts, fixed seeds and deterministic training options, containerized runtime (Docker/Singularity), versioned datasets, and published checkpoints together with clear evaluation pipelines and pre-specified metrics. Ethical deployment calls for routine bias and robustness audits, privacy-preserving data practices, human-in-the-loop decision workflows, and prospective validation across diverse cohorts and acquisition settings before clinical use. Together these steps reduce the risk of overfitting to simulation artifacts and improve trustworthiness in real-world settings.

Practically, ensemble XAI can accelerate clinical prediction, vaccine-response modeling, autoimmune monitoring, biomarker discovery, and safer preclinical screeningprovided predictions are accompanied by concise, verifiable explanations and governance artifacts that clinicians can inspect. Future work must focus on multi-center benchmarking, experimental validation of top-ranked features, and integration with clinical workflows under regulatory and ethical oversight.

References

[1]. Athaya T, Ripan RC, Li X, Hu H. Multimodal deep learning approaches for single-cell multi-omics data integration. Brief Bioinform. 2023 Sep 20;24(5):bbad313. doi: 10.1093/bib/bbad313. PMID: 37651607; PMCID: PMC10516349.Fabris, A. (2023). *Fairness and Bias in Algorithmic Hiring* (survey). arXiv 2023. https://arxiv.org/pdf/2309.13933

[2]. Jin W, Yang Q, Chi H, Wei K, Zhang P, Zhao G, Chen S, Xia Z, Li X. Ensemble deep learning enhanced with self-attention for predicting immunotherapeutic responses to cancers. Front Immunol. 2022 Dec 1; 13:1025330. doi: 10.3389/fimmu.2022.1025330. PMID: 36532083; PMCID: PMC9751999.

[3]. A. Maslova, R.N. Ramirez, K. Ma, H. Schmutz, C. Wang, C. Fox, B. Ng, C. Benoist, S. Mostafavi, & Immunological Genome Project, Deep learning of immune cell differentiation, Proc. Natl. Acad. Sci. U.S.A. 117 (41) 25655-25666, https://doi.org/10.1073/pnas.2011795117 (2020).

[4]. Yang M et al. Multimodal data deep learning method for predicting symptomatic pneumonitis in combined radiotherapy and immunotherapy. Frontiers in Immunology. 2025

[5]. Maria Hügle, Gabriel Kalweit, Thomas Huegle, Joschka Boedecker A Dynamic Deep Neural Network For Multimodal Clinical Data AnalysisarXiv:2008.06294 [cs.LG](or arXiv:2008.06294v1 [cs.LG] for this version)https://doi.org/10.48550/arXiv.2008.06294

[6]. Abanades, B., Wong, W.K., Boyles, F. et al. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. Commun Biol 6, 575 (2023). https://doi.org/10.1038/s42003-023-04927-7

[7]. Seung Min Baik, Kyung Sook Hong, Jae-Myeong Lee, Dong Jin Park, integrating ensemble and machine learning models for early prediction of pneumonia mortality using laboratory tests,Heliyon, Volume 10, Issue 14, 2024, e34525, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2024.e34525.

[8]. Montesinos-López OA, Chavira-Flores M, Kiasmiantini, Crespo-Herrera L, Saint Piere C, Li H, Fritsche-Neto R, Al-Nowibet K, Montesinos-López A, Crossa J. A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding. Genetics. 2024 Nov 5;228(4): iyae161. doi: 10.1093/genetics/iyae161. Epub ahead of print. Erratum in: Genetics. 2025 Feb 5;229(2): iyae200. doi: 10.1093/genetics/iyae200. PMID: 39499217; PMCID: PMC11631469.

[9]. Chia-Ru Chung, Chung-Yu Chien, Yun Tang, Li-Ching Wu, Justin Bo-Kai Hsu, Jang-Jih Lu, Tzong-Yi Lee, Chen Bai, Jorng-Tzong Horng, An ensemble deep learning model for predicting minimum inhibitory concentrations of antimicrobial peptides against pathogenic bacteria, iScience,Volume 27, Issue 9,2024, 110718, ISSN 2589-0042, https://doi.org/10.1016/j.isci.2024.110718.

[10]. Ing, A., Andrades, A., Cosenza, M.R. et al. Integrating multimodal cancer data using deep latent variable path modelling. Nat Mach Intell 7, 1053–1075 (2025). https://doi.org/10.1038/s42256-025-01052-4

[11]. Zhenjiang Fan, Jie Sun, Henry Thorpe, Stephen Lee, Soyeon Kim, Hyun Jung Park, Deep neural network learning biological condition information refines gene-expression-based cell subtypes, Briefings in Bioinformatics, Volume 25, Issue 1, January 2024, bbad512, https://doi.org/10.1093/bib/bbad512

[12]. Btissam Bouzammour, Ghita Zaz, Malika Alami Marktani, Hiba Chougrad, Abdellah Touhafi, Mohammed Jorio, Ensemble deep learning models for respiratory disease detection using cough analysis, Journal of Engineering Research,2025, ISSN 2307-1877, https://doi.org/10.1016/j.jer.2025.08.017.