Privacy-Preserving Deep Neural Inference for Clinical Prediction Using Adaptive Polynomial Activation Approximation and Encrypted Medical Data

Shannan Kennelly
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
shannan1999@ucf.edu

Wout van Salm
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
woutmail@ku.edu

Matteo Dippel
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
dippelmatteo@unr.edu

Abstract

Clinical prediction models increasingly rely on deep neural networks trained on electronic health records, laboratory values, vital signs, and imaging-derived features. Although such models can capture nonlinear interactions relevant to risk stratification, cloud-hosted or third-party inference exposes sensitive medical data and intermediate representations. Homomorphic encryption offers a principled way to evaluate a network on encrypted inputs, but its algebraic limitations make nonlinear activations difficult to compute. Fixed polynomial replacements can work on a nominal interval, yet they may fail when activation distributions shift across patient subgroups, clinical sites, or layers. This manuscript develops a privacy-preserving inference framework that couples encrypted medical data with adaptive polynomial activation approximation. The method selects layer-specific approximation intervals, polynomial degrees, and encrypted normalization parameters by minimizing distribution-weighted approximation error under a noise-depth budget. It then propagates approximation error through the network to obtain a risk-score perturbation bound and links that bound to clinical decision margins and abstention rules. The analysis clarifies how adaptive degree allocation differs from uniform polynomial replacement, why activation-range normalization is important for deep encrypted inference, and how leakage boundaries arise from ciphertext access patterns, approximation statistics, and model architecture. The contribution is methodological and analytical: it provides a coherent design for encrypted clinical prediction, derives explicit trade-offs among activation fidelity, multiplicative depth, and decision stability, and identifies failure modes that must be controlled before such models can be used in privacy-sensitive clinical workflows.

Keywords

homomorphic encryption; clinical prediction; polynomial activation approximation; encrypted neural inference; noise-aware planning; decision calibration

1. Introduction

Homomorphic-encryption-based neural inference has progressed from proof-of-concept evaluation of shallow networks on encrypted inputs [1] to approximate arithmetic that supports real-valued matrices and deeper computation [2]. Layer-wise polynomial evaluation [3], privacy-preserving classification [4], encrypted deep networks [5], and low-latency designs [6] have established that useful neural computation can be performed without exposing plaintext inputs to the evaluating party. Clinical prediction is a natural target because the inputs are often high-dimensional, personally identifiable, and distributed across institutions. Deep models trained on electronic health records can integrate longitudinal laboratory measurements, vital signs, diagnoses, and utilization histories, and they have been studied as scalable alternatives to manually engineered risk scores [7]. Architectural advances have made deep learning on electronic health records more practical [8], and surveys of deep electronic health record modeling document both the promise and the methodological hazards of such models [9]. The central technical obstacle is not the linear algebra of a dense layer but the handling of nonlinear activation functions under encryption, because homomorphic ciphertexts support additions and multiplications while activations such as rectified linear units, sigmoids, and Gaussian error linear units are not directly compatible with those operations.

Privacy risk in clinical machine learning is not hypothetical. Model inversion can recover representative features from confidence information [10], membership inference can determine whether a record was used in training [11], and white-box inference attacks can extract information from gradients or parameters under unfavorable assumptions [12]. Defenses such as private training through knowledge transfer [13] reduce some leakage during model development, but they do not remove the need for protected inference when a patient record or a local hospital data source is sent to an external computation environment. Homomorphic encryption addresses a different part of the problem: the server evaluates a model while the input, intermediate values, and often the output remain encrypted, and the client decrypts only the final result. This separation is attractive for clinical prediction because the evaluating server need not observe laboratory values, diagnosis codes, or derived activations.

The central difficulty is that encrypted inference for a deep clinical model requires a numerical approximation of every nonlinearity, and the approximation must be chosen before the server can inspect the encrypted data. A polynomial of fixed degree and fixed interval may be adequate for a nominal activation distribution, but clinical data are heterogeneous. A model applied to intensive care unit admissions, outpatient laboratory panels, or specialty-specific registries can generate different preactivation ranges and different tail behavior. A degree that is too low introduces systematic bias and can move a patient across a decision threshold; a degree that is too high consumes multiplicative depth, increases noise, and may force bootstrapping or a less secure parameter set. Uniform activation approximation therefore forces an undesirable compromise: it spends computational resources evenly even though only some layers dominate output sensitivity, and it applies one interval even though activation distributions differ across layers and subgroups. This mismatch is the design gap addressed here.

The proposed framework is an adaptive polynomial activation approximation method for privacy-preserving deep neural inference on encrypted medical data. The method has four connected parts. First, it defines layer-wise approximation error with respect to an activation

distribution informed by clinical cohort statistics, rather than a generic uniform interval. Second, it selects polynomial degrees and approximation intervals by solving a constrained resource-allocation problem in which approximation error competes with multiplicative depth and ciphertext noise. Third, it introduces encrypted normalization parameters so that a fixed normalized polynomial can be evaluated over a range aligned to the layer's activation scale without revealing patient-specific values to the server. Fourth, it propagates approximation error through the network to derive a bound on risk-score perturbation and connects that bound to clinical decision margins, abstention rules, and review pathways. The analysis is deliberately analytical rather than empirical: it derives design consequences from the algebraic structure of encrypted inference, polynomial approximation theory, and sensitivity of clinical decisions, without attributing unperformed measurements or unobserved performance to any dataset.

The contribution is a coherent inference design rather than a library of isolated approximations. It explains why adaptive degree allocation is preferable to uniform replacement, how normalization changes the approximation problem, why noise growth must be treated as part of activation selection, and where privacy leakage can persist despite encryption. The discussion also identifies boundary conditions: activation intervals must cover clinically plausible tails, approximation statistics must be aggregated or protected, and decision margins must reflect both numerical error and calibration uncertainty. These issues matter because a privacy-preserving clinical model that is cryptographically sound but numerically fragile can still produce unstable risk estimates.

2. Background and Related Work

2.1 Encrypted neural inference

The first generation of encrypted neural inference demonstrated that a small number of layers could be evaluated by replacing nonlinearities with low-degree polynomials and packing values into ciphertext slots [1]. Approximate-number homomorphic encryption subsequently made it practical to represent real-valued activations and weights directly, at the cost of approximate arithmetic and explicit scale management [2]. Discretized network evaluation introduced sign and step approximations that map naturally to certain ciphertext operations, but those approximations alter gradients and can be unsuitable when a clinical model's output is used probabilistically [3]. Other designs examined privacy-preserving classification with deep networks [4] and encrypted convolutional or fully connected models with specialized packing [5]. Low-latency inference work showed that scheduling, data layout, and activation replacement can matter as much as the underlying cryptographic primitive [6]. These studies motivate the present focus on activation approximation, but they generally assume a fixed approximation strategy for each nonlinearity.

Cryptographic backends impose different constraints. Torus-based schemes provide exact gate-level operations and programmable bootstrapping, which can realize arbitrary functions through look-up-table-like operations, but they require substantial recoding of real-valued clinical computations [14]. Approximate-number schemes support matrix arithmetic more naturally but require careful management of multiplicative depth and scale. Polynomial evaluation theory supplies the bridge between these worlds: Horner evaluation and Paterson-Stockmeyer schedules trade multiplications, storage, and depth in ways that directly affect the feasibility of encrypted inference [15]. Specialized encrypted network systems have exploited sparsity, batching, and activation-aware schedules to reduce latency [16], and secure logistic regression studies have shown that clinical risk models can be evaluated under encryption

with careful numerical design [17]. The framework developed here abstracts away from a single backend: it specifies activation approximation, normalization, and depth allocation in a backend-neutral form, then leaves the concrete polynomial evaluation schedule to the chosen homomorphic library.

2.2 Clinical prediction and privacy risk

Clinical prediction models often combine heterogeneous inputs: continuous laboratory values with skewed distributions, missingness indicators, categorical diagnosis codes, medication histories, and time-varying vital signs. Deep models can learn interactions among these variables, but their outputs are used in decisions that are sensitive to small probability changes, particularly when a risk score sits near a treatment threshold. The privacy concern is therefore twofold. The input record is sensitive, and the model's intermediate representations may also reveal information about the patient or the training population. Inversion and membership attacks demonstrate that confidence values and model behavior can leak more than intended [10, 11]. White-box attacks show that parameter access can amplify leakage under some training and deployment conditions [12], while private training methods reduce leakage during model fitting but do not encrypt inference [13]. Encrypted inference addresses the deployment side of the problem by keeping the patient record encrypted while the server performs the model evaluation. It does not, by itself, provide differential privacy or protect against all side channels, so its guarantees must be stated precisely.

2.3 Activation approximation gap

Activation approximation is often treated as an implementation detail: replace rectified linear units with a square, a low-degree minimax polynomial, or a piecewise approximation, then evaluate the rest of the network normally. This approach can be adequate for benchmark networks, but it leaves three questions unresolved. First, what interval should the polynomial approximate? A minimax polynomial over a wide interval sacrifices accuracy in the high-probability region, whereas a narrow interval clips clinically plausible tails. Second, how should degree vary across layers? A layer whose outputs strongly influence the final risk may need a higher degree than a layer whose contribution is attenuated by subsequent small weights. Third, how should numerical error be connected to the clinical decision rule? An approximation that is accurate in mean squared error may still shift a patient across a threshold if the error is biased in the threshold region.

Automatic ReLU replacement for private inference treats activation substitution as a search problem and seeks fast private network evaluation by replacing nonlinear units with HE-compatible alternatives [18]. That line of work is relevant because it recognizes that activation choice is a system-level decision rather than a fixed preprocessing step. The present framework differs in emphasis. It does not search over arbitrary activation replacements for speed alone; it formulates a distribution-aware approximation problem whose objective includes clinical decision stability, encrypted normalization, and noise-depth allocation. The two perspectives are complementary: search can identify candidate replacements, while the present analysis explains how to choose approximation intervals and degrees when the downstream quantity is a calibrated clinical risk.

3. Problem Formulation and Threat Model

Let x denote a patient feature vector after local preprocessing, including imputation, standardization, and missingness indicators. Let W_l and b_l denote the weights and biases of layer l, and let σ denote the activation function. A deep network with L layers computes a

scalar or vector risk score by alternating affine transformations and nonlinear activations. In encrypted inference, the client holds x in plaintext, encrypts it, and sends ciphertexts to a server that holds the model parameters. The server returns an encrypted score to the client.

Here f_l denotes the operation of layer l, including its affine transformation and activation, y is the network output, and the ellipsis indicates composition across layers. The server should not learn x, any intermediate activation, or y. The client should learn y but not necessarily the model weights if the model is proprietary. The threat model considered here is an honest-but-curious server that follows the protocol but may inspect ciphertexts, access patterns, and protocol metadata. The client may also be curious about the model if weights are encrypted, but the primary clinical privacy concern is protection of the patient record and its derived representations. The framework assumes that the underlying homomorphic encryption scheme provides semantic security for ciphertexts and that key management is handled by the client. It does not claim protection against a server that observes decryption results, compromises the client, or exploits unencrypted side channels unless those channels are explicitly controlled.

Clinical prediction introduces requirements that generic encrypted inference benchmarks often omit. The output must remain usable as a probability or risk score after approximation. The decision rule may involve a threshold τ, a calibrated risk band, or a triage rule that refers uncertain cases for human review. Missingness patterns and outlier laboratory values are clinically meaningful and should not be collapsed into an undifferentiated numerical range. The model may be updated across sites, so approximation coefficients must be revisable without exposing patient-level data. These requirements motivate an inference design that separates three quantities: the encrypted input, the public or encrypted normalization parameters, and the polynomial coefficients used for activation approximation.

The proposed inference protocol assumes that polynomial coefficients are selected before inference using aggregate activation statistics from a development cohort or a privacy-protected aggregation process. At inference time, the server evaluates a fixed polynomial schedule on encrypted activations. If patient-specific normalization is used, the client encrypts the normalization parameters together with the input, or the server derives them from encrypted summaries using a protocol that does not decrypt them. The server never sees the raw activation distribution. This separation is important because activation statistics from training data can reveal information about clinical subgroups if published at fine granularity. Coefficients should therefore be derived from aggregated, de-identified, or noise-protected summaries, or from conservative physiological ranges when such summaries are unavailable.

4. Methodological Framework

4.1 Distribution-aware polynomial activation approximation

For each layer k, let z_k denote the preactivation vector and let μ_k denote the activation distribution induced by the development cohort or a clinically representative reference population. The target activation σ is approximated by a polynomial p_k of degree d_k over an interval A_k. The approximation error is measured as a weighted L2 norm with respect to μ_k, not as a uniform error over an arbitrary interval.

$$\varepsilon_k(d_k, A_k) \;=\; (\int_\{A_k\} (\sigma(z) \;-\; p_k(z))^2 \, d\mu_k(z))\text{^}\{1/2\}$$

In this expression, ε_k is the layer-k approximation error, d_k is the polynomial degree, A_k is the approximation interval, σ is the activation function, p_k is the polynomial

approximation, and μ_k is the activation distribution. The interval A_k is chosen to contain a high-probability portion of μ_k, with an explicit tail reserve for clinical outliers. The polynomial is represented in a numerically stable basis for coefficient computation and then converted to a monomial or Chebyshev form for homomorphic evaluation.

$$p_k(z) = \sum_{j=0}^{d_k} c_{k,j} z^j$$

Here c_{k,j} are the polynomial coefficients for layer k. The coefficients can be obtained by weighted least squares or a weighted minimax procedure on A_k, with a small regularization term to control ill-conditioning when d_k is large. The use of μ_k is what makes the approximation adaptive: the same activation function can produce different effective intervals in an intensive care model, an outpatient risk model, and a readmission model because the preactivation distributions differ. A polynomial that is excellent on a nominal interval can be poor on a shifted interval, and a polynomial that is fitted to one subgroup can be unstable for another.

The interval A_k can be parameterized by a center m_k and a scale s_k, for example as a symmetric range around a robust center or as a quantile-bounded interval. The center and scale need not be patient identifiers; they are normalization parameters. When these parameters are public, the server can evaluate a normalized polynomial directly. When they are patient-specific, they are encrypted and treated as part of the ciphertext computation.

4.2 Encrypted normalization and fixed normalized polynomials

A central design choice is to separate adaptation from evaluation. Instead of choosing a new polynomial for every patient at inference time, the framework normalizes the preactivation into a canonical interval and applies a polynomial q_k that has been fitted to that normalized interval.

$$\tilde{z}_k = (z_k - m_k)/s_k, p_k(z_k) = q_k(\tilde{z}_k)$$

In this equation, z̃_k is the normalized preactivation, m_k is the center parameter, s_k is the scale parameter, and q_k is the normalized polynomial. The server computes z̃_k homomorphically using encrypted z_k and encrypted or public m_k and s_k. The polynomial q_k is then evaluated on the normalized variable. This separation has two benefits. It allows the approximation error to be controlled on a stable interval, and it allows the same polynomial coefficients to be reused across patients while the normalization parameters absorb shifts in activation scale. If m_k and s_k are encrypted, the server does not learn the patient's activation center or scale. If they are public cohort-level parameters, they reveal only aggregate information.

The normalization step also controls clipping bias. If the normalized activation falls outside the approximation interval, the polynomial may diverge. The client or the model designer must therefore specify a clipping range that reflects clinically plausible outliers. Clipping is not merely a numerical convenience: it changes the effective activation and should be included in the approximation error budget. In practice, the interval A_k should be selected so that the weighted error ε_k includes both the polynomial fit error inside the interval and a conservative estimate of the tail mass outside it.

4.3 Noise-aware degree and interval allocation

Homomorphic polynomial evaluation consumes multiplicative depth. A high-degree polynomial can approximate a difficult activation more accurately, but it also increases ciphertext noise and may require bootstrapping or a larger ciphertext modulus. The framework therefore treats degree selection as a resource-allocation problem. For planning purposes, let D(d) denote the multiplicative depth cost of evaluating a degree-d polynomial under a balanced Paterson-Stockmeyer-style schedule. A normalized abstraction is sufficient for design reasoning.

$$D(d) = \lceil log_2(d + 1) \rceil$$

Here D(d) is the multiplicative depth cost, d is the polynomial degree, and the ceiling function reflects the discrete number of depth levels. The exact constant depends on the evaluation schedule, the homomorphic library, and whether powers are reused across layers. The expression is used as a planning abstraction rather than as a claim about a specific implementation. Given a total depth budget D_max, the adaptive planner selects degrees and intervals to minimize a combination of approximation error and depth cost.

$$min \sum_\{k = 1\}\^\{L\} [\varepsilon_k(d_k, A_k) + \lambda D(d_k)] \; s.t. \sum_\{k = 1\}\^\{L\} D(d_k) \leq D_max$$

In this optimization, ε_k(d_k, A_k) is the layer-k approximation error from the distribution-weighted criterion, D(d_k) is the depth cost, λ is a positive trade-off parameter, D_max is the maximum multiplicative depth available under the chosen cryptographic parameters, and L is the number of layers. The minimization is over the degree d_k and interval A_k for each layer. The Lagrange multiplier λ can be interpreted as the marginal cost of depth in units of approximation error. A layer with large downstream sensitivity receives a higher degree or a better interval because its ε_k is weighted more heavily in the propagated error bound. A layer whose effect is attenuated by small subsequent weights can use a lower degree.

This optimization is the principal difference between adaptive polynomial activation approximation and uniform replacement. Uniform replacement assigns the same degree to every layer or chooses degree based on the worst-case layer, which wastes depth. Adaptive allocation concentrates numerical accuracy where it changes the output. It also allows the interval A_k to be selected jointly with the degree: a narrower interval with a moderate degree can outperform a wider interval with a high degree when the activation distribution is concentrated. The planner must account for clinical outlier risk, so the interval should not be so narrow that tail patients are systematically distorted.

4.4 Encrypted clinical inference protocol

The protocol has five stages. In the first stage, the client performs local preprocessing: imputation, standardization using site-specific parameters, encoding of missingness indicators, and any bounded transformation that reduces extreme values. These steps occur in plaintext on the client side and are part of the clinical data pipeline. In the second stage, the client encrypts the feature vector and, if used, the normalization parameters. In the third stage, the server evaluates the linear layers using homomorphic additions and multiplications by plaintext weights, followed by the normalized polynomial activation for each layer. If the model weights must also be protected, the linear layers can be evaluated with encrypted weights, but that increases cost and changes the depth budget. The fourth stage, bootstrapping, is invoked only when the noise budget approaches its limit; the adaptive planner tries to avoid bootstrapping by allocating depth across layers. In the fifth stage, the client decrypts the risk score and applies the decision rule.

The server's evaluation schedule should be fixed or padded so that timing does not reveal which layers used larger polynomials or which branches were taken. Although the ciphertext hides the patient values, protocol metadata can still reveal model architecture, the number of layers, and the frequency of bootstrapping. In a clinical setting, such metadata may be less sensitive than the record itself but should not be treated as automatically safe. The framework therefore recommends a deterministic evaluation schedule with coefficients and degrees fixed before the inference request. Patient-specific normalization parameters, if used, should be encrypted with the same scheme and packed in a way that avoids revealing their position or magnitude.

4.5 Clinical decision calibration

Encrypted inference is useful only if the decrypted risk score supports a defensible clinical decision. Approximation error can shift a risk score across a threshold, so the decision rule should use an error margin. Let r(x) be the exact model risk and $\hat{r}(x)$ be the approximate risk produced by the encrypted polynomial network. Suppose the propagated error bound satisfies $|r(x) - \hat{r}(x)| \leq E$ for the relevant input domain. A threshold decision is stable when the approximate risk is sufficiently far from the threshold.

$$|\hat{r}(x) - \tau| > E \Rightarrow \textit{decision stable under approximation error } E$$

In this expression, $\hat{r}(x)$ is the approximate risk, τ is the clinical decision threshold, and E is the propagated error bound. If the approximate risk lies within the band $[\tau - E, \tau + E]$, the decision is not guaranteed to match the exact model. The clinical workflow can then request human review, use a higher-precision polynomial schedule if the depth budget permits, or apply a conservative default rule. This is not a cryptographic guarantee about patient outcomes; it is a numerical stability condition that makes the approximation error explicit at the point of decision. The margin should be communicated to clinicians as part of model validation, and it should be recalibrated when the model, cohort, or encoding pipeline changes.

5. Analytical Assessment

5.1 Approximation error propagation

The layer-wise error ε_k measures how well p_k approximates σ on the activation distribution, but the final risk depends on how those errors propagate through subsequent layers. Suppose the activation approximation at layer k introduces an error bounded by ε_k on the relevant interval, and suppose the activation and affine transformations are Lipschitz with constants that can be bounded from the model weights. A first-order propagation bound can be written as follows.

$$||r - \hat{r}|| \leq \sum_{k=1}^{L} \left(\prod_{i=k+1}^{L} L_i \, ||W_i||\right) \varepsilon_k$$

Here r is the exact network output, $\hat{r}$ is the approximate output, L_i is the Lipschitz constant associated with layer i's activation on the relevant interval, W_i is the weight matrix, and ε_k is the approximation error at layer k. The product term accounts for amplification of an error introduced at layer k by all later layers. The bound is conservative because it uses worst-case Lipschitz constants and does not exploit cancellation, but it provides a useful design principle: approximation accuracy should be allocated according to downstream amplification. A layer followed by large weight norms and expansive activations deserves a smaller ε_k than a layer whose output is compressed by later transformations.

The bound also exposes a failure mode that uniform approximation hides. If a deep network has large weight norms or repeated expansive layers, a small local approximation

error can become a large output perturbation. In that case, increasing the polynomial degree in the first layer alone may not be sufficient; the planner must consider the sensitivity of the entire path. The distribution-weighted error ε_k is preferable to a uniform error because it penalizes errors where the activation mass actually lies. However, it does not eliminate the need for a tail reserve. A patient with an extreme laboratory value can fall outside the high-probability interval even if the cohort-weighted error is small. Clinical deployment should therefore include input-range checks and a fallback mode for values outside the validated domain.

5.2 Depth, noise, and computational trade-offs

The adaptive allocation problem couples numerical accuracy to cryptographic feasibility. Increasing d_k reduces ε_k but raises $D(d_k)$, which consumes multiplicative depth and increases noise. In approximate-number homomorphic encryption, each multiplication also affects scale, so polynomial evaluation must be scheduled with rescaling and, when necessary, bootstrapping. The depth abstraction $D(d) = \lceil \log_2(d + 1) \rceil$ captures the qualitative trade-off: doubling the degree does not necessarily double depth, but it does increase the number of ciphertext multiplications and the noise accumulated in the polynomial evaluation. A planner that ignores noise may choose high degrees that are mathematically attractive but cryptographically impractical. A planner that ignores approximation error may choose low degrees that are fast but clinically unstable.

The adaptive method improves on fixed-degree allocation in three ways. First, it can assign a higher degree to layers with large downstream amplification and a lower degree to layers whose errors are attenuated. Second, it can enlarge or shift the approximation interval in layers where the activation distribution is heavy-tailed, reducing clipping bias without forcing every layer to use a wide interval. Third, it can choose normalization parameters that make the normalized activation more concentrated, which lowers the degree required to achieve a target ε_k. These benefits are analytical consequences of the optimization; they do not rely on a particular dataset or benchmark result.

There is a countervailing cost. Adaptive allocation requires activation statistics from a development cohort, and those statistics must be representative of the deployment population. If the deployment site has a different laboratory assay, coding practice, or patient mix, the chosen intervals may be misaligned. A conservative planner should use broader intervals or site-specific recalibration when distribution shift is plausible. The depth budget should also include a reserve for bootstrapping or for a higher-precision fallback when a patient's encrypted input falls outside the validated range. This reserve reduces the risk that a rare but clinically important outlier is evaluated with a polynomial that is only valid near the center of the distribution.

5.3 Privacy boundaries

Homomorphic encryption protects plaintext values under the semantic security of the scheme, but it does not make all information invisible. The server observes ciphertext sizes, the number of layers, the polynomial degrees, the frequency of rotations, and the timing of the evaluation. If the protocol uses different degrees for different patients, the access pattern may reveal something about the patient's activation range or the model's confidence. The framework mitigates this by fixing the evaluation schedule before the request and by using encrypted normalization parameters when patient-specific scaling is required. It does not

claim that these measures eliminate all side channels; they reduce the most direct leakage paths.

Aggregate activation statistics used to fit the polynomials are another potential leakage source. If coefficients are published at fine granularity for small subgroups, they may reveal distributional information about those subgroups. The planner should therefore use sufficiently aggregated statistics, coarse quantization, or a privacy mechanism when the cohort is small. The polynomial coefficients themselves are less sensitive than patient records, but they are not automatically public. In a multi-site deployment, coefficients can be computed through secure aggregation or derived from a public reference population and then recalibrated locally without sharing patient-level data. The clinical decision margin E should also be treated as model metadata that may reveal sensitivity information; it need not be hidden from clinicians, but it should not be exposed to an untrusted server if the threat model includes model confidentiality.

6. Technical Discussion

6.1 Relation to alternative privacy-preserving approaches

The proposed framework sits between generic encrypted inference and activation-specific replacement search. Compared with secure multi-party computation, homomorphic encryption avoids interactive communication during the inference computation, which is attractive for asynchronous clinical workflows. The cost is higher per-operation computation and a more constrained nonlinear evaluation. Compared with trusted execution environments, homomorphic encryption reduces trust in the hardware vendor but shifts complexity to cryptographic parameter selection and numerical approximation. Compared with federated learning, encrypted inference protects the input at prediction time rather than the training data during model fitting; the two approaches address different parts of the privacy lifecycle. Compared with differential privacy, homomorphic encryption hides individual values from the server but does not by itself bound information leakage from aggregate model releases. These distinctions matter in clinical deployments because privacy claims are often conflated.

Within encrypted inference, the closest alternative to adaptive polynomial activation is a fixed activation replacement chosen for speed. Automatic ReLU replacement for private inference demonstrates that search over HE-compatible activations can improve private network evaluation [18]. The present framework is complementary but asks a different question: given a clinical decision threshold and a depth budget, how should approximation intervals and polynomial degrees be allocated so that the decrypted risk remains stable? This question cannot be answered by speed alone because a faster approximation that biases the threshold region may be clinically unacceptable. It also cannot be answered by accuracy alone because a high-degree approximation may consume depth that the cryptographic parameters cannot support. The adaptive planner makes both considerations explicit.

6.2 Failure modes and boundary conditions

The first failure mode is interval misspecification. If the development cohort underrepresents a clinical subgroup, the polynomial may be fitted to an interval that is too narrow for that subgroup. The result is clipping bias or polynomial divergence outside the interval. Mitigation requires conservative interval selection, subgroup-aware validation, and input-range monitoring. The second failure mode is error amplification in deep networks with large weight norms. The propagation bound shows that a small local error can become a large output error if subsequent layers are expansive. Regularization during model training and

sensitivity-aware degree allocation can reduce this risk, but they do not eliminate it. The third failure mode is calibration drift. A model can remain numerically accurate while its risk estimates become miscalibrated after changes in clinical practice or documentation. Approximation error bounds do not replace calibration monitoring; they add a numerical margin that must be combined with clinical validation.

A fourth failure mode is noise-budget exhaustion. If the planner underestimates the depth required for polynomial evaluation, the server may need to bootstrap more often or use a larger ciphertext modulus. This can change latency, increase the attack surface through timing, or force a security parameter trade-off. The depth abstraction should be calibrated against the chosen homomorphic library during implementation, and the planner should include a safety margin. A fifth failure mode is leakage through metadata. Fixed schedules and padded batching reduce but do not eliminate this risk. A clinical deployment should treat the inference protocol as part of the security architecture, not as an implementation detail hidden behind a cryptographic library.

6.3 Implications for clinical prediction systems

The framework suggests that privacy-preserving clinical prediction should be designed around the decision rule, not only around the network architecture. A model that is accurate on average but unstable near a treatment threshold is a poor candidate for encrypted inference with approximate activations unless the approximation error is bounded and the workflow provides a review path for uncertain cases. The adaptive activation method gives model developers a way to translate a depth budget into a numerical accuracy budget and then into a decision margin. This translation is useful for regulatory review because it makes explicit which numerical assumptions support the encrypted evaluation. It also gives clinical sites a way to reason about distribution shift: if a new site’s patient population differs, the activation intervals and decision margins should be re-evaluated.

The approach does not eliminate the need for careful data governance. Encrypted inference protects the record during computation, but the client still controls decryption, the server still observes metadata, and the model provider still controls coefficients and parameters. Clinical deployment requires key management, audit logging, and a clear separation of responsibilities between the site, the model provider, and the computation provider. These are operational concerns, but they are inseparable from the technical design because a numerically fragile or metadata-leaky protocol can undermine the privacy benefits of encryption even when the underlying ciphertext operations are secure.

7. Conclusion

This manuscript developed a privacy-preserving inference framework for clinical prediction that combines adaptive polynomial activation approximation with encrypted medical data. The core problem is that homomorphic encryption can evaluate linear operations efficiently but cannot directly evaluate the nonlinear activations on which deep clinical models depend. Fixed polynomial replacements are insufficient because activation distributions vary across layers, patient subgroups, and clinical sites, and because approximation error interacts with clinical decision thresholds. The proposed framework addresses this problem by selecting layer-specific intervals and polynomial degrees through a distribution-weighted approximation criterion, by normalizing activations under encryption, and by allocating multiplicative depth according to downstream sensitivity. The resulting

design connects approximation fidelity to risk-score perturbation and to a clinical decision margin that can trigger review or a higher-precision fallback.

The analysis yields three defensible conclusions. First, activation approximation should be treated as a resource-allocation problem rather than a uniform preprocessing choice. Higher-degree polynomials are valuable only when their error reduction matters for the final risk and when the cryptographic depth budget can support them. Second, encrypted normalization is a practical way to reconcile a fixed normalized polynomial with patient- or site-specific activation scales without revealing those scales to the server. Third, numerical error must be propagated to the decision rule. A privacy-preserving model that is cryptographically secure but numerically unstable near a threshold is not suitable for clinical use. The framework also identifies privacy boundaries that encryption alone does not remove, including metadata leakage and aggregate approximation statistics. By making these trade-offs explicit, the proposed approach provides a technically grounded basis for designing encrypted clinical prediction systems that are both privacy-preserving and decision-aware.

References

1. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016). CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the 33rd International Conference on Machine Learning (PMLR 48, pp. 201–210).

2. Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic encryption for arithmetic of approximate numbers. In Advances in Cryptology – ASIACRYPT 2017 (LNCS 10624, pp. 409–437). Springer.

3. Bourse, F., Minelli, M., Minihold, M., & Paillier, P. (2018). Fast homomorphic evaluation of deep discretized neural networks. In Advances in Cryptology – CRYPTO 2018 (LNCS 10993, pp. 483–512). Springer.

4. Chabanne, H., de Wargny, A., Milgram, J., Morel, C., & Prouff, E. (2017). Privacy-preserving classification on deep neural network. Cryptology ePrint Archive, Report 2017/035.

5. Hesamifard, E., Takabi, H., & Ghasemi, M. (2017). CryptoDL: Deep neural networks over encrypted data. arXiv:1711.05189.

6. Brutzkus, A., Gilad-Bachrach, R., & Erez, S. (2019). Low latency privacy preserving inference. In Proceedings of the 36th International Conference on Machine Learning (PMLR 97, pp. 812–821).

7. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094.

8. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Duggan, G. E., Flores, G., Laramie, D., Li, Y., Long, X., Lu, J., … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1, 18.

9. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604.

10. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322–1333). ACM.

11. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (pp. 3–18). IEEE.

12. Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In 2019 IEEE Symposium on Security and Privacy (pp. 739–753). IEEE.

13. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In International Conference on Learning Representations.

14. Chillotti, I., Gama, N., Georgieva, M., & Izabachène, M. (2020). TFHE: Fast fully homomorphic encryption over the torus. Journal of Cryptology, 33(1), 34–91.

15. Paterson, M. S., & Stockmeyer, L. J. (1973). On the number of nonscalar multiplications necessary to evaluate polynomials. SIAM Journal on Computing, 2(1), 60–66.

16. Lou, Q., & Jiang, L. (2019). SHE: A fast and accurate deep neural network for encrypted data. In Advances in Neural Information Processing Systems 32.

17. Kim, M., Song, Y., Wang, S., Xia, Y., & Jiang, X. (2018). Secure logistic regression based on homomorphic encryption: Design and evaluation. JMIR Medical Informatics, 6(2), e19.

18. Peng, H., Huang, S., Zhou, T., Luo, Y., Wang, C., Wang, Z., ... & Ding, C. (2023, October). Autorep: Automatic relu replacement for fast private network inference. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 5155-5165). IEEE.