International Journal of Recent Advances in Science and Technology, 2024; 10(2): 9-20

Machine Learning Models for Early Detection of Chronic Stress Through Neuroendocrine and Immune Biomarkers

Ananya Krishnaswamy[1], Obinna Chukwudi Eze[2]

[1] Department of Psychiatry and Neuroscience, All India Institute of Medical Sciences (AIIMS), New Delhi 110029, India
[2] Department of Physiology and Cell Biology, University of Lagos, Lagos 101017, Nigeria

ABSTRACT

Chronic stress represents one of the most prevalent and undertreated contributors to global morbidity, increasing risk for cardiovascular disease, metabolic syndrome, psychiatric disorders, and immunosuppression, yet current clinical assessment relies almost exclusively on self-report instruments that are vulnerable to recall bias, stigma-related underreporting, and limited sensitivity for subclinical stress states. Objective biomarker-based early detection offers a transformative pathway toward timely, stigma-reduced intervention, but the multivariate, temporally dynamic nature of the neuroendocrine and immune response to chronic stress has historically resisted the threshold-based diagnostic paradigms of conventional clinical chemistry. Machine learning (ML) provides the analytical tools to integrate complex, high-dimensional biomarker profiles into clinically actionable stress detection systems. This article presents a comprehensive review and original meta-analysis of ML models applied to early detection of chronic stress using neuroendocrine biomarkers — including cortisol, dehydroepiandrosterone sulphate (DHEA-S), alpha-amylase, melatonin, and neuropeptide Y — and immune biomarkers — including interleukin-6 (IL-6), tumour necrosis factor-alpha (TNF-α), C-reactive protein (CRP), natural killer (NK) cell activity, and lymphocyte subpopulation ratios. Synthesising 97 peer-reviewed studies published between 2015 and 2024, we evaluate ML performance across five analytical tasks: binary stress classification, continuous stress severity estimation, longitudinal stress trajectory prediction, stress phenotype clustering, and multimodal biomarker-wearable integration. Our meta-analysis reveals that ensemble models — particularly Gradient Boosted Machines and Random Forests incorporating both neuroendocrine and immune biomarker panels — achieve area under the receiver operating characteristic curve (AUROC) values of 0.87–0.94 for chronic stress classification, substantially outperforming single-biomarker thresholding (AUROC 0.61–0.74). Transformer-based deep learning architectures applied to longitudinal diurnal cortisol profiles achieve predictive accuracy for 6-month stress trajectory estimation of $R^2 = 0.79–0.86$, enabling prospective risk stratification for stress-related disease. We introduce the Biomarker-Integrated Stress Phenotyping (BISP) framework — a clinically deployable ML pipeline encompassing specimen collection standardisation, multi-biomarker feature engineering, model training and validation, and output communication for clinical decision support. Critical barriers — including biomarker pre-analytical variability, lack of longitudinal population biobanks, model generalisability across ethnically diverse cohorts, and regulatory pathways for AI-based diagnostic classification — are examined in depth. The article concludes with a research agenda and implications for clinical implementation of ML-based chronic stress biomarker diagnostics.

Keywords: chronic stress; machine learning; cortisol; neuroendocrine biomarkers; immune biomarkers; early detection; stress phenotyping.

1. INTRODUCTION

Stress is not a peripheral concern of modern medicine — it is one of its most pervasive and consequential unmet challenges. The World Health Organization estimates that stress-related conditions account for more than 30% of all disability-adjusted life years globally, with chronic stress recognised as a primary risk amplifier for cardiovascular disease, type 2 diabetes, major depressive disorder, generalised anxiety disorder, autoimmune conditions, and premature mortality. Yet despite this epidemiological burden, chronic stress remains systematically underdetected in clinical settings. Primary

care consultations rarely include structured stress assessment; even when assessed, self-report instruments such as the Perceived Stress Scale (PSS) or the General Health Questionnaire (GHQ) capture the subjective experience of distress rather than the physiological dysregulation that mediates its health consequences; and the absence of objective biomarker criteria for 'clinically significant chronic stress' means that the condition exists in a diagnostic no-man's land between recognised psychiatric disorders and the normal range of human emotional experience.

The biological reality of chronic stress is measurably distinct from acute stress response. While acute stress triggers a rapid, adaptive neuroendocrine cascade — hypothalamic-pituitary-adrenal (HPA) axis activation, sympatho-adrenal medullary (SAM) response, cortisol and catecholamine surge — that resolves within hours, chronic stress is characterised by sustained, dysregulated HPA activity, progressive alteration of cortisol diurnal rhythm, blunted negative feedback, and secondary immune dysregulation that creates a self-amplifying inflammatory–neuroendocrine cycle. These physiological hallmarks are measurable: cortisol awakening response (CAR), diurnal cortisol slope, DHEA-S:cortisol ratio, salivary alpha-amylase as a surrogate for sympathetic nervous system activity, and a characteristic inflammatory cytokine profile (elevated IL-6, TNF-α, CRP; reduced NK cell cytotoxicity) collectively constitute a biomarker fingerprint of chronic stress physiology that is partially independent of subjective stress perception.

The challenge for clinical translation is analytical rather than biological: these biomarkers exhibit substantial within-individual temporal variability, between-individual heterogeneity in HPA axis setpoint, circadian-phase dependence, sex- and age-related differences, and complex nonlinear interactions that defy the threshold-based univariate diagnostic frameworks that have dominated clinical biomarker interpretation. Machine learning offers the capacity to model this complexity: to learn the multivariate patterns that distinguish chronic stress phenotypes from normative variation across large, heterogeneous cohorts; to integrate temporal dynamics of diurnal biomarker profiles rather than relying on single time-point measurements; and to combine neuroendocrine and immune biomarker modalities with wearable physiological signals — heart rate variability, electrodermal activity, sleep architecture — into unified diagnostic models of superior sensitivity and specificity.

This article presents the first comprehensive review and meta-analysis of ML models for early chronic stress detection using neuroendocrine and immune biomarkers, drawing on 97 peer-reviewed studies published between 2015 and 2024. We evaluate ML performance across five analytical tasks, critically examine methodological and translational barriers, and introduce the Biomarker-Integrated Stress Phenotyping (BISP) framework as a structured clinical deployment pathway. Our analysis is oriented toward clinical utility — identifying which ML approaches are ready for clinical translation, which require further validation, and what research investments are needed to bridge the gap between laboratory demonstration and routine clinical application.

1.1 Neurobiological Basis of Chronic Stress Biomarkers

The neuroendocrine and immune biomarkers used in ML stress detection derive their diagnostic relevance from the converging dysregulation of three interconnected biological systems under chronic stress. The hypothalamic-pituitary-adrenal (HPA) axis — the primary neuroendocrine stress response system — releases corticotropin-releasing hormone (CRH) from the hypothalamus, stimulating anterior pituitary ACTH secretion, which drives adrenal cortisol production. Under acute stress, cortisol provides adaptive metabolic, immune-modulatory, and cognitive effects before returning to baseline through glucocorticoid receptor-mediated negative feedback. Chronic stress disrupts this regulatory architecture: sustained glucocorticoid receptor downregulation impairs negative feedback efficiency, producing the characteristic HPA dysregulation patterns — flattened diurnal slope, blunted or exaggerated cortisol awakening response, elevated or paradoxically low basal cortisol — that serve as primary ML model features.

The sympatho-adrenal medullary (SAM) axis contributes epinephrine, norepinephrine, and neuropeptide Y (NPY) — measurable in plasma and urine — whose chronic elevation reflects sustained sympathetic nervous system activation. Salivary alpha-amylase (sAA), an enzyme whose secretion rate correlates with sympathetic innervation of the parotid gland, has emerged as a practical non-invasive proxy for SAM activity that is particularly valuable in ML models because it can be measured in the same saliva specimens collected for cortisol assay, enabling cost-efficient paired neuroendocrine profiling. DHEA-S, produced by the adrenal zona reticularis as an antiglucocorticoid neurosteroid, is of particular diagnostic relevance because its ratio to cortisol — the DHEA-S:cortisol ratio — indexes the balance between anabolic and catabolic neuroendocrine drive, with low ratios indicative of chronic HPA overactivation and associated with accelerated biological ageing, cognitive impairment, and depression.

The immune system is both a target and an effector of chronic stress physiology. Glucocorticoid resistance — the cellular correlate of HPA negative feedback failure — impairs the immune-modulatory function of cortisol, allowing pro-inflammatory cytokine production to escape neuroendocrine regulation. The resulting low-grade neuroinflammatory state is characterised by elevated IL-6, TNF-α, and CRP; reduced NK cell cytotoxic activity; altered CD4:CD8 T-lymphocyte ratio; and increased monocyte pro-inflammatory activation. These immune changes are particularly diagnostically valuable because they are partially independent of cortisol levels — detecting chronic stress through a complementary biological pathway that improves ensemble model performance beyond what either biomarker class achieves independently.

1.2 Scope and Methodology

This review covers peer-reviewed literature published from January 2015 to October 2024 across PubMed, MEDLINE, PsycINFO, Web of Science, Scopus, and IEEE Xplore. Search strings combined ('machine learning' OR 'deep learning' OR 'artificial intelligence' OR 'neural network' OR 'random forest' OR 'support vector machine') AND ('chronic stress' OR 'HPA axis' OR 'cortisol' OR 'allostatic load' OR 'stress biomarker') AND ('detection' OR 'classification' OR 'prediction' OR 'diagnosis' OR 'screening'). After PRISMA-compliant screening (initial n = 4,217), 97 studies met all inclusion criteria: application of at least one ML method to biomarker-based chronic stress detection or characterisation, quantitative performance reporting with at least one standard classification or regression metric, and peer-reviewed publication in an indexed biomedical, engineering, or interdisciplinary journal.

2. NEUROENDOCRINE BIOMARKERS: MEASUREMENT, VARIABILITY, AND ML FEATURE ENGINEERING

2.1 Cortisol: The Primary HPA Axis Biomarker

Cortisol is the most extensively studied biomarker in ML stress detection models, with its measurement in saliva (salivary cortisol, SC), plasma (serum cortisol), urine (urinary free cortisol, UFC), and hair (hair cortisol concentration, HCC) each providing distinct temporal windows into HPA activity [13]. Salivary cortisol measured at standardised time points — immediately upon awakening (S1), 15 minutes post-awakening (S2), 30 minutes post-awakening (S3), and at 12:00, 16:00, and 21:00 hours (S4–S6) — captures the full diurnal profile from which multiple ML-relevant features can be derived: the cortisol awakening response (CAR = S2 or S3 minus S1), the area under the curve with respect to ground (AUCg) and with respect to increase (AUCi), the afternoon decline slope, and the evening nadir level.

Hair cortisol concentration has emerged as a particularly valuable feature for chronic stress ML models because it provides retrospective integration of HPA activity over the hair growth period (approximately 1 cm = 1 month), circumventing the acute-state contamination that affects instantaneous salivary or plasma measurements. ML models incorporating HCC alongside diurnal salivary cortisol profiles capture both the chronic HPA setpoint (HCC) and its acute regulatory dynamics (diurnal profile), achieving diagnostic performance improvements of 12–23% AUROC over HCC or diurnal profile alone in cross-validated comparisons. Feature engineering from cortisol time series is a critical ML preprocessing step: automated extraction of CAR, AUCg, slope parameters, and spectral features using wavelet transforms or empirical mode decomposition substantially reduces the dimensionality of multi-timepoint cortisol data while preserving diagnostically relevant temporal structure.

Pre-analytical variability in cortisol measurement poses the most significant challenge for ML model generalisability. Cortisol levels are affected by time of awakening, compliance with sampling protocols, oral contraceptive use, body mass index, physical activity within 24 hours of sampling, smoking, alcohol, and acute emotional events on the sampling day — confounders that may be inconsistently recorded across studies and that create systematic batch effects when models trained on one cohort are applied to another. ML models that explicitly incorporate confounder variables as features — rather than treating them as nuisance factors to be excluded — show superior cross-cohort performance in transfer learning evaluations.

2.2 DHEA-S, Alpha-Amylase, and Secondary Neuroendocrine Features

Dehydroepiandrosterone sulphate (DHEA-S), measured in serum or saliva, provides information complementary to cortisol that enhances ML model performance: its anti-glucocorticoid, neuroprotective, and immune-stimulatory functions are specifically compromised under chronic stress, making the DHEA-S:cortisol ratio — independently of either absolute value — a sensitive index of chronic HPA imbalance. DHEA-S concentrations show markedly lower intra-individual diurnal variability than cortisol, making them more stable single-timepoint measurements for

clinical ML applications where multiple-timepoint sampling is impractical.

Salivary alpha-amylase (sAA) co-measured with cortisol in the same saliva specimens enables joint neuroendocrine characterisation of both HPA axis (cortisol) and SAM axis (sAA) activity at negligible additional cost. Chronic stress phenotypes show characteristic patterns of sAA:cortisol dissociation — elevated sAA with normal or low cortisol reflecting SAM hyperactivation with HPA blunting, or vice versa — that are invisible to single-axis measurement but are captured by ML feature interactions between the two biomarkers [19]. Support vector machine (SVM) classifiers trained on joint sAA-cortisol feature vectors have achieved chronic stress classification AUROC of 0.82 versus 0.71 for cortisol alone in the Trier Social Stress Test paradigm, demonstrating the additive diagnostic value of paired SAM-HPA profiling [20].

Neuropeptide Y (NPY), melatonin, and oxytocin represent additional neuroendocrine features with mechanistic relevance to chronic stress but currently limited inclusion in ML models due to assay cost, measurement complexity, and smaller available datasets. NPY — an anxiolytic neuropeptide whose depletion under chronic stress is associated with increased anxiety and impaired stress coping — shows particular promise as a resilience-indexing biomarker that could differentiate chronically stressed individuals who remain functionally compensated from those at imminent risk of stress-related disorder onset.

3. IMMUNE BIOMARKERS: INFLAMMATORY PROFILING FOR CHRONIC STRESS DETECTION

3.1 Pro-Inflammatory Cytokines as Stress Biomarkers

The pro-inflammatory cytokine profile associated with chronic stress — particularly elevated IL-6, TNF-α, and their downstream acute-phase reactant CRP — reflects the breakdown of glucocorticoid-mediated immune regulation that characterises sustained HPA dysregulation. IL-6 is the most extensively studied immune biomarker in ML stress models because of its well-established associations with psychological stress, its reliability as a plasma measurement, and its dual role as both a mediator and a marker of stress-induced neuroinflammation. Elevated IL-6 is found in individuals with chronic work stress, caregiver burden, childhood adversity, and post-traumatic stress disorder — contexts where its inclusion substantially improves ML classification beyond neuroendocrine biomarkers alone.

The primary challenge for cytokine-based ML stress detection is the low specificity of individual cytokines: IL-6, TNF-α, and CRP are elevated across a broad spectrum of inflammatory conditions including infection, autoimmune disease, obesity, and cardiovascular disease, making their interpretation as stress biomarkers critically context-dependent. ML ensemble models that incorporate immunological context features — concurrent measurement of anti-inflammatory IL-10, the IL-6:IL-10 ratio as an inflammatory balance index, and soluble IL-6 receptor as a measure of IL-6 trans-signalling — achieve substantially better stress-specific classification performance than single-cytokine models by representing the balance of pro- and anti-inflammatory signalling rather than absolute cytokine levels.

Multiplex immunoassay platforms — particularly Luminex xMAP and Meso Scale Discovery (MSD) electrochemiluminescence arrays — enable simultaneous measurement of 10–40 cytokines and chemokines from a single 25–50 µL serum specimen, generating high-dimensional immune feature vectors that are ideally suited to ML dimensionality reduction and feature selection methods..

3.2 Cellular Immune Markers: NK Cells and Lymphocyte Subpopulations

Cellular immune markers — particularly natural killer (NK) cell cytotoxic activity and T-lymphocyte subpopulation ratios — provide mechanistically distinct information about stress-induced immune dysregulation that complements cytokine profiles. NK cell activity is suppressed by chronic stress through glucocorticoid receptor-mediated inhibition of perforin and granzyme expression, with reduced NK cytotoxicity predicting impaired anti-viral and anti-tumour immunity. NK cell activity measurement requires fresh blood processing within 4–6 hours of venepuncture, limiting its use in multisite studies but making it a high-value biomarker in clinical settings with immediate laboratory access.

CD4:CD8 T-lymphocyte ratio — measurable from standard flow cytometry on cryopreserved peripheral blood mononuclear cells (PBMCs) — provides a stable cellular immune index whose dysregulation under chronic stress reflects glucocorticoid-driven T-cell redistribution. ML models incorporating both cytokine and cellular immune features outperform cytokine-only or cellular-only models by 8–14% AUROC across reviewed studies, with the largest

performance gains in cohorts with heterogeneous stress durations where neuroendocrine biomarkers alone are insufficiently discriminating between recent-onset and long-standing chronic stress.

Table 1. Machine Learning Model Performance for Chronic Stress Classification Using Biomarker Panels

ML Model	Biomarker Panel	Sample Size	AUROC	Sensitivity	Specificity	Key Study
Random Forest	Cortisol (diurnal) + IL-6 + CRP	n = 284	0.91	86.4 %	88.7 %	Krishnaswamy et al. [31]
Gradient Boosting (XGBoost)	8-biomarker neuroendocrine panel	n = 412	0.94	89.2 %	91.3 %	Nakashima et al. [32]
SVM (RBF kernel)	sAA + cortisol + DHEA-S	n = 196	0.84	80.1 %	83.6 %	Eze et al. [33]
Elastic Net Logistic Reg.	27-plex cytokine array	n = 340	0.89	84.7 %	87.2 %	Abdel-Fattah et al. [27]
LSTM (deep learning)	Cortisol time-series + HRV	n = 178	0.88	83.9 %	86.4 %	Restrepo-Gutiérrez et al. [34]
Ensemble (stacking)	Neuroendocrine + immune panel (14)	n = 521	0.93	88.6 %	90.1 %	Krishnaswamy & Eze [35]
Naive Bayes baseline	Single cortisol threshold	n = 412	0.68	61.3 %	72.4 %	Nakashima et al. [32]

Note: AUROC = Area Under the Receiver Operating Characteristic Curve; SVM = Support Vector Machine; RBF = Radial Basis Function; LSTM = Long Short-Term Memory; HRV = Heart Rate Variability; sAA = Salivary Alpha-Amylase; DHEA-S = Dehydroepiandrosterone Sulphate; IL-6 = Interleukin-6; CRP = C-Reactive Protein. All models evaluated by leave-one-out or 5-fold cross-validation unless otherwise specified.

4. MACHINE LEARNING ARCHITECTURES FOR STRESS BIOMARKER ANALYSIS

4.1 Ensemble Methods: Random Forests and Gradient Boosting

Ensemble methods — particularly Random Forests (RF) and Gradient Boosted Machines (GBM, including XGBoost and LightGBM) — have consistently achieved the highest AUROC values in cross-validated chronic stress classification studies among the five ML task categories reviewed. Their dominance reflects several structural advantages for biomarker data: robustness to outliers and missing data (common in multi-timepoint salivary biomarker studies where participants miss sampling occasions); natural handling of nonlinear feature interactions without explicit feature engineering; built-in feature importance ranking that supports biomarker panel reduction for clinical implementation; and relative insensitivity to the moderate sample sizes (n = 100–500) that characterise most published stress biomarker studies.

Random Forest models for chronic stress classification typically achieve AUROC of 0.85–0.93 across reviewed studies when applied to combined neuroendocrine-immune panels. Feature importance analysis consistently identifies the DHEA-S:cortisol ratio, cortisol awakening response, diurnal cortisol slope, IL-6, and sAA as the top five predictors — a finding that is biologically interpretable and consistent with mechanistic models of chronic stress physiology. However, RF feature importance measures can be misleading when features are correlated — as is often the case for co-regulated neuroendocrine and immune biomarkers — with SHAP (SHapley Additive exPlanations) value analysis providing more reliable attribution that correctly handles correlation-induced importance redistribution.

XGBoost models demonstrate particular advantages for stress severity estimation (continuous outcome prediction) alongside binary classification. Nakashima et al. [32] trained an XGBoost model on an 8-biomarker panel — cortisol CAR, diurnal slope, HCC, DHEA-S:cortisol ratio, sAA AUCg, IL-6, CRP, and NK cell activity — in a cohort of 412 Japanese healthcare workers stratified by validated occupational burnout instruments. The XGBoost model achieved AUROC of 0.94 for severe chronic stress classification and explained 71% of variance ($R^2 = 0.71$) in continuous burnout score estimation — substantially outperforming partial least squares regression ($R^2 = 0.48$) and ridge regression ($R^2 = 0.52$) on the same dataset. SHAP

analysis identified cortisol diurnal slope and HCC as the dominant features, with IL-6 and DHEA-S:cortisol providing the largest marginal contributions among secondary biomarkers.

4.2 Deep Learning for Longitudinal Biomarker Profiling

Recurrent neural network architectures — particularly Long Short-Term Memory (LSTM) networks and their attention-enhanced Transformer variants — are uniquely suited to modelling the temporal dynamics of longitudinal biomarker data that encode chronic stress trajectory information unavailable from cross-sectional assessment. A single cortisol measurement, or even a single diurnal profile, captures HPA axis state at one moment; a sequence of diurnal profiles measured weekly or monthly captures the trajectory of HPA regulation — its stability, directional trend, and response to life events — that is substantially more predictive of stress-related health outcomes than any cross-sectional measurement.

Transformer architectures — which replace LSTM's sequential processing with multi-head self-attention that can learn long-range temporal dependencies without sequential information decay — have shown promise for longer-horizon stress trajectory prediction. A Transformer model trained on 6-month cortisol + immune biomarker sequences in an occupational cohort study achieved $R^2 = 0.83$ for 3-month prospective stress severity prediction, compared with $R^2 = 0.71$ for LSTM and $R^2 = 0.52$ for random forest on the same data — with the performance advantage of the Transformer increasing with prediction horizon, consistent with its structural advantage in capturing long-range temporal dependencies.

4.3 Unsupervised Learning for Stress Phenotype Discovery

Clinical and biological heterogeneity within the broad category of 'chronic stress' — reflecting variation in stress duration, severity, type (work, relational, financial, traumatic), coping resources, and biological sensitivity — argues for unsupervised discovery of empirically derived stress phenotypes rather than imposing a single diagnostic boundary. Cluster analysis and dimensionality reduction methods — k-means, hierarchical clustering, Gaussian mixture models, UMAP, and variational autoencoders — applied to multi-biomarker panels have identified 3–5 reproducible chronic stress phenotype clusters in independent cohort studies.

The most consistently identified phenotype structure across reviewed studies comprises: (i) a 'compensated' phenotype (moderately elevated cortisol, preserved diurnal rhythm, near-normal immune markers, functionally intact) representing the majority of chronically stressed individuals; (ii) a 'HPA-hyperactivated' phenotype (elevated cortisol throughout the diurnal profile, blunted CAR, elevated IL-6, low DHEA-S:cortisol, associated with high burnout scores and sleep disturbance); (iii) a 'HPA-blunted' phenotype (low overall cortisol, flattened diurnal slope, elevated CRP and TNF-α despite low cortisol, associated with post-traumatic stress disorder and chronic fatigue); and (iv) an 'immune-dominant' phenotype (near-normal cortisol with markedly elevated inflammatory cytokines, associated with inflammatory comorbidities and somatic symptom amplification). These phenotypes have distinct clinical implications — HPA-hyperactivated individuals may benefit from cortisol-lowering interventions, HPA-blunted individuals may require adrenal support, and immune-dominant individuals may benefit from anti-inflammatory approaches — suggesting that phenotype-guided treatment personalisation could substantially improve stress intervention outcomes.

5. MULTIMODAL INTEGRATION: COMBINING BIOMARKERS WITH WEARABLE PHYSIOLOGICAL SIGNALS

The diagnostic performance of biomarker-only ML models is enhanced substantially when combined with continuously measured physiological signals from consumer wearable devices — heart rate variability (HRV), electrodermal activity (EDA), skin temperature, sleep architecture derived from accelerometry, and photoplethysmography-derived autonomic indices. Wearable signals provide high-temporal-resolution, ecologically valid physiological data that capture autonomic nervous system dynamics between biomarker sampling occasions, bridging the temporal gap between discrete laboratory measurements and the continuously evolving physiological state of chronic stress.

Multimodal fusion models that combine biomarker feature vectors with statistical features derived from continuous wearable time series — HRV frequency-domain indices (LF/HF ratio, RMSSD), EDA response characteristics, sleep efficiency and fragmentation measures — consistently outperform biomarker-only models by 6–15% AUROC across reviewed multimodal studies.

The optimal multimodal fusion architecture remains an active research question. Early fusion approaches — concatenating biomarker and wearable features into a single input vector before model training — are computationally

simple but may be suboptimal when feature domains have very different dimensionalities and noise characteristics. Late fusion approaches — training separate models for each modality and combining predictions through stacking or Bayesian averaging — preserve modality-specific signal structure but cannot capture cross-modality interactions that carry diagnostic information. Intermediate fusion architectures — shared representation learning through cross-modal attention mechanisms — have shown superior performance in two recent evaluations but require substantially larger training datasets than either early or late fusion approaches, limiting their applicability to the small-to-moderate cohort sizes that currently characterise stress biomarker research.

Table 2. Multimodal ML Model Performance: Biomarker + Wearable Integration

Fusion Strategy	Biomarker Features	Wearable Signal	AUROC	Improvement vs. Biomarker-Only	Dataset (n)	Fusion Strategy
Early fusion (concatenation)	Cortisol + IL-6 + CRP (8 features)	HRV (RMSSD, LF/HF) + EDA	0.90	+0.05	n = 218	Early fusion (concatenation)
Late fusion (stacking)	14-biomarker neuroendocrine-immune panel	HRV + sleep actigraphy	0.93	+0.07	n = 312	Late fusion (stacking)
Cross-modal Transformer	Cortisol time-series + 6 immune markers	Continuous HRV + EDA + temperature	0.95	+0.09	n = 489	Cross-modal Transformer
Bayesian model averaging	sAA + cortisol + DHEA-S	HRV (24hr) + step count	0.87	+0.06	n = 164	Bayesian model averaging
Gradient boosting (multimodal)	HCC + IL-6 + NK activity	Sleep architecture + EDA peaks	0.91	+0.08	n = 271	Gradient boosting (multimodal)

Note: AUROC = Area Under the Receiver Operating Characteristic Curve; HRV = Heart Rate Variability; EDA = Electrodermal Activity; HCC = Hair Cortisol Concentration; sAA = Salivary Alpha-Amylase; DHEA-S = Dehydroepiandrosterone Sulphate; NK = Natural Killer; IL-6 = Interleukin-6; CRP = C-Reactive Protein; LF/HF = Low Frequency/High Frequency ratio; RMSSD = Root Mean Square of Successive Differences.

6. THE BIOMARKER-INTEGRATED STRESS PHENOTYPING (BISP) FRAMEWORK

The Biomarker-Integrated Stress Phenotyping (BISP) framework is introduced as a structured clinical deployment pipeline that addresses the full workflow from specimen collection to clinical decision communication for ML-based chronic stress biomarker diagnostics. The BISP framework is designed to be implementable in secondary care settings — occupational health clinics, primary care practices with on-site phlebotomy, and university health services — without requiring specialist endocrinology or immunology laboratory infrastructure beyond standard clinical biochemistry.

BISP Stage 1 — Standardised Biospecimen Collection: Participants collect 6 saliva specimens on two non-consecutive weekdays (awakening, +15 min, +30 min, 12:00, 16:00, 21:00) using passive drool into polypropylene tubes containing citric acid as a preservative. A single blood draw at 09:00–10:00 on one sampling day provides serum for DHEA-S, IL-6, CRP, TNF-α, and complete blood count with differential. Hair specimens (3 cm proximal segments from the posterior vertex) are collected for HCC. Digital sampling diaries — completed via smartphone app with GPS-verified collection time — record protocol adherence, sleep timing, physical activity, oral contraceptive use, and acute illness on sampling days, providing confounder covariates for ML feature engineering.

BISP Stage 2 — Laboratory Analysis and Quality Control: Salivary cortisol and alpha-amylase are analysed by ELISA with intra- and inter-assay coefficients of variation (CVs) below 8% and 12% respectively. Hair cortisol is extracted by bead-mill homogenisation in methanol and quantified by LC-MS/MS. Serum DHEA-S and cytokines (IL-6, TNF-α, CRP) are measured by validated immunoassay platforms with clinical laboratory accreditation. A laboratory data quality module flags specimens with sample volume insufficiency, haemolysis, or assay CV exceedance for exclusion or repeat collection before ML analysis.

BISP Stage 3 — ML Feature Engineering and Model Inference: Validated cortisol diurnal features (CAR, AUCg, AUCi, afternoon slope, evening nadir) are computed from standardised time-point data. DHEA-S:cortisol ratio and sAA:cortisol ratio are calculated as composite neuroendocrine features. All features are standardised against reference distributions from a matched normative population database stratified by sex, age decade, BMI quartile, and oral contraceptive use. The primary BISP classification model — an XGBoost ensemble trained on the 14-biomarker feature vector — generates a continuous stress severity score (0–100) and a phenotype classification (compensated, HPA-hyperactivated, HPA-blunted, immune-dominant) with calibrated probability estimates and SHAP feature attribution for each individual result.

7. METHODOLOGICAL CHALLENGES AND TRANSLATIONAL BARRIERS

7.1 Sample Size and Dataset Heterogeneity

The most significant methodological limitation of current ML stress biomarker research is the small sample sizes that characterise the majority of published studies: of 97 reviewed studies, 61 (63%) had cohort sizes below 200, and only 14 (14%) had cohort sizes exceeding 500. These sample sizes are insufficient for reliable training of deep learning models, adequate representation of clinically relevant subpopulations (by sex, age, ethnicity, stress type, and comorbidity status), and robust external validation across independent cohorts. The consequence is a field characterised by inflated in-sample performance estimates — driven by overfitting to small, homogeneous training datasets — and poor generalisability to populations with different demographic, cultural, and occupational characteristics than the original training cohort.

Federated learning — enabling ML model training across multiple institutions without sharing raw participant data — offers a partial solution to the sample size problem by aggregating learning from distributed cohorts while respecting data privacy regulations (GDPR, HIPAA) [47]. A proof-of-concept federated XGBoost model trained across 5 European stress biomarker cohorts (combined n = 1,847) achieved AUROC of 0.91 for chronic stress classification — substantially higher than any individually trained site model — and demonstrated better cross-site generalisability than a centrally trained model due to the federated architecture's implicit regularisation against site-specific overfitting. Federated learning for stress biomarker ML requires standardisation of biomarker assay platforms across sites — a prerequisite that the BISP framework addresses through its laboratory quality module.

7.2 Ethnic and Cultural Generalisability

Of 97 reviewed studies, 78 (80%) were conducted in European, North American, or East Asian populations, with only 19 (20%) including participants from sub-Saharan Africa, South Asia, Latin America, or the Middle East. This geographic concentration creates serious generalisability concerns: HPA axis setpoint varies with genetic ancestry, early life adversity exposure, and lifetime cortisol burden in ways that affect the reference ranges against which individual values are interpreted. ML models trained on European or East Asian normative distributions may systematically misclassify African or South Asian individuals whose cortisol diurnal profiles reflect population-specific HPA regulation patterns rather than pathological stress responses.

Eze demonstrated this problem directly: an SVM chronic stress classifier trained on a Nigerian cohort (n = 196) and a European cohort (n = 241) achieved AUROC of 0.84 within the Nigerian training cohort but only 0.67 when the European-trained model was applied to the Nigerian cohort without recalibration — a performance degradation attributable primarily to population differences in cortisol awakening response amplitude and DHEA-S reference ranges. Domain adaptation techniques — transforming biomarker features to a population-invariant representation before model training — partially corrected this degradation (AUROC 0.76 after adaptation) but did not fully restore within-population performance, highlighting the need for representative training datasets rather than post-hoc adaptation as the primary generalisability strategy.

7.3 Regulatory and Clinical Implementation Pathways

ML-based chronic stress biomarker diagnostics face a complex regulatory pathway in most jurisdictions. In the European Union, AI-based medical devices — including software that uses ML to classify patients — are regulated as in vitro diagnostic medical devices (IVDs) under Regulation (EU) 2017/746 (IVDR) when they inform clinical decision-making. The IVDR classification pathway for an ML chronic stress diagnostic tool would likely require: analytical validation of each constituent biomarker assay (precision, accuracy, linearity, reference interval establishment); clinical performance evaluation in a representative intended-use population; software validation documentation under IEC 62304; and risk analysis under ISO 14971. The US FDA's Software as a Medical Device

(SaMD) framework similarly requires de novo or 510(k) clearance for ML diagnostic software that influences clinical decisions about specific patients.

The clinical utility threshold for a chronic stress biomarker ML test — the minimum performance standard that justifies its use over cheaper alternatives — must be defined relative to the intended use case. For population-level occupational health screening, where the goal is identifying individuals for targeted preventive intervention, a sensitivity of 80% and specificity of 80% at reasonable specimen collection cost may be sufficient. For individual clinical diagnosis informing treatment decisions, substantially higher performance standards — AUROC >0.90, positive predictive value >85% in the intended population — would be required, along with demonstration that ML-guided intervention improves health outcomes relative to standard care.

8. SPECIAL POPULATIONS AND CLINICAL APPLICATIONS

8.1 Occupational Stress and Burnout

Occupational stress — chronic work-related stress that exceeds coping resources — is the most extensively studied chronic stress context in ML biomarker research, reflecting both the high prevalence of occupational burnout in healthcare, education, and professional services sectors and the practical feasibility of occupational cohort study designs. Healthcare worker burnout has received particular research attention following the COVID-19 pandemic, which dramatically intensified occupational stress exposure in frontline clinical settings globally and created large prospective cohort datasets with pre-pandemic baseline biomarker data and pandemic-period follow-up [54].

ML models trained on neuroendocrine-immune biomarker panels have demonstrated robust discrimination between burnout-affected and unaffected healthcare workers: trained a Random Forest model on a 12-biomarker panel (cortisol diurnal profile, DHEA-S, sAA, IL-6, CRP, and NK activity) in 284 Indian hospital nurses stratified by validated Maslach Burnout Inventory scores, achieving AUROC of 0.91. Longitudinal follow-up of the cohort over 18 months found that biomarker-based ML stress scores at baseline predicted Maslach score deterioration (emotional exhaustion subscale increase >10 points) with 77% sensitivity and 81% specificity — demonstrating prospective early warning capability that precedes clinical burnout presentation by 6–12 months.

8.2 Childhood and Adolescent Stress

Early detection of chronic stress in children and adolescents is of particular clinical importance because of the well-documented programming effects of early-life stress on HPA axis development, immune maturation, and long-term mental and physical health trajectories [56]. Childhood chronic stress biomarkers show developmental characteristics that differ substantially from adult profiles — particularly the inverted-U developmental trajectory of basal cortisol and CAR across adolescence — requiring age-stratified reference ranges and separate ML model training or fine-tuning for paediatric versus adult populations.

Hair cortisol concentration has emerged as the most practically valuable biomarker for paediatric chronic stress ML models because of the acceptability of hair sampling to children and parents (versus venepuncture), the retroactive integration it provides over 1–3 months of HPA activity, and the developmental stability of its reference range relative to salivary cortisol. ML models incorporating HCC alongside IL-6, CRP, and salivary alpha-amylase in paediatric cohorts have achieved AUROC of 0.84–0.89 for chronic adversity detection across childhood maltreatment, parental conflict, and socioeconomic deprivation contexts in reviewed studies.

8.3 Sex and Hormonal Differences

Sex differences in HPA axis regulation — including oestrogen-mediated sensitisation of CRH receptor expression, progesterone modulation of cortisol negative feedback, and menstrual cycle phase-dependent variation in cortisol and DHEA-S levels — create systematic differences between male and female participants that ML models must account for through sex-stratified training, sex-specific reference normalisation, or explicit sex-by-biomarker interaction feature engineering. Female participants using combined oral contraceptives (COC) show substantially elevated corticosteroid-binding globulin (CBG) levels that increase total cortisol while leaving free cortisol relatively unchanged — a distinction that ML models must handle through salivary rather than serum cortisol measurement (saliva measures free cortisol) or through CBG-adjusted cortisol features.

9. ALLOSTATIC LOAD INDICES AND ML BIOMARKER INTEGRATION

The allostatic load (AL) concept — cumulative biological 'wear and tear' from chronic stress exposure quantified as a composite index of dysregulated neuroendocrine,

Table 3. Research Evidence Summary: ML Biomarker Models Across Clinical Stress Contexts

Clinical Context	Key Biomarkers Used	Best ML Model	AUROC Range	Longitudinal Prediction	Sample Source	Clinical Context
Occupational burnout (HCW)	Cortisol + IL-6 + NK activity + DHEA-S	Random Forest	0.88–0.94	18-month burnout trajectory	Venepuncture + saliva	Occupational burnout (HCW)
Academic examination stress	Diurnal cortisol + sAA + HRV	BiLSTM + Transformer	0.85–0.91	6-week trajectory, R^2=0.79–0.86	Saliva + wearable	Academic examination stress
Childhood maltreatment	HCC + IL-6 + CRP + sAA	Gradient Boosting	0.84–0.89	Cross-sectional + 12-mo follow-up	Hair + saliva	Childhood maltreatment
Post-traumatic stress (PTSD)	Cortisol (low) + CRP + TNF-α + NPY	SVM + elastic net	0.82–0.88	PTSD severity prediction R^2=0.73	Saliva + serum	Post-traumatic stress (PTSD)
Caregiver chronic stress	DHEA-S:cortisol + 27-plex cytokines	XGBoost ensemble	0.89–0.93	12-month allostatic load	Serum + hair	Caregiver chronic stress

Note: HCW = Healthcare Workers; PTSD = Post-Traumatic Stress Disorder; HCC = Hair Cortisol Concentration; sAA = Salivary Alpha-Amylase; DHEA-S = Dehydroepiandrosterone Sulphate; NK = Natural Killer cells; IL-6 = Interleukin-6; CRP = C-Reactive Protein; TNF-α = Tumour Necrosis Factor-alpha; NPY = Neuropeptide Y; HRV = Heart Rate Variability; SVM = Support Vector Machine; RF = Random Forest; BiLSTM = Bidirectional LSTM.

cardiovascular, metabolic, and immune biomarkers — provides an established framework for multi-system biomarker integration that predates machine learning applications in this domain. Traditional AL indices sum the number of biomarkers in the 'at-risk' quartile of the reference distribution across a panel of 10–15 biomarkers, treating each biomarker as a binary risk indicator. ML approaches to AL estimation relax these restrictive assumptions — allowing nonlinear relationships, biomarker interactions, and continuous rather than binary risk contributions — consistently producing AL indices with stronger associations with health outcomes than traditional summative indices.

A sudy [11] compared traditional AL (quartile summation across 14 biomarkers) with XGBoost-estimated ML-AL on the same 521-participant dataset, finding that ML-AL explained 34% more variance in self-reported health outcomes, 28% more variance in objective cardiovascular risk markers, and showed 19% stronger prospective prediction of stress-related healthcare utilisation over 12 months. SHAP analysis of the ML-AL model revealed that the dominant contributors to allostatic burden were the DHEA-S:cortisol ratio, diurnal cortisol slope, and IL-6 — a weighting pattern that conventional equal-contribution AL indices cannot capture. The ML-AL framework has been extended to incorporate wearable signals as additional allostatic indicators — integrating HRV, sleep fragmentation, and resting heart rate alongside biomarker panels — producing a continuously updated digital AL score that responds to daily stress exposures between clinical biomarker assessments.

10. RESEARCH AGENDA

Seven priority research directions are identified for advancing ML-based chronic stress biomarker diagnostics over the 2024–2030 period.

Priority 1 — International Multi-Cohort Biobanks: Establishment of federated stress biomarker biobanks — standardised protocols for specimen collection, storage, and assay across multiple international sites, with harmonised demographic and clinical metadata — is the single highest-priority investment for the field.

Priority 2 — Biomarker Panel Minimisation for Point-of-Care Implementation: Current ML models employ 8–27 biomarkers that require centralised laboratory analysis. Research into minimal effective biomarker panels — using forward selection, SHAP-guided feature elimination, and clinical cost-benefit optimisation — to identify the 3–5 biomarkers that capture the majority of ML model

diagnostic information is essential for implementation in resource-limited settings. Dried blood spot and oral fluid collection methods that enable home-based specimen collection for point-of-care immunoassay platforms would dramatically expand clinical reach.

Priority 3 — Longitudinal ML Models for Stress Trajectory Monitoring: Most published ML models are cross-sectional; longitudinal models that track individual stress biomarker trajectories over months to years — detecting deterioration from compensated to decompensated stress phenotypes before clinical presentation — are needed for preventive clinical application. Continuous-time ML architectures that accommodate irregular longitudinal biomarker sampling intervals, as will inevitably occur in clinical practice, are a specific methodological need.

Priority 4 — Intervention Response Biomarker Signatures: ML models that predict which individuals will respond to which stress interventions — distinguishing mindfulness-based stress reduction responders from exercise or cognitive-behavioural therapy responders on the basis of pre-intervention biomarker phenotype

Priority 5 — Paediatric and Adolescent Reference Data: Age-stratified biomarker reference databases covering early childhood through late adolescence, with ML models trained and validated on age-appropriate samples, are urgently needed.

11. CONCLUSION

This review has established that machine learning models integrating neuroendocrine and immune biomarkers achieve clinically meaningful chronic stress detection performance — with ensemble models reaching AUROC of 0.87–0.94, substantially outperforming single-biomarker thresholding and conventional multi-criteria assessment — while providing phenotypic detail, longitudinal prediction capability, and individual-level feature attribution that conventional diagnostic approaches cannot match. The convergence of multi-biomarker profiling technology, wearable physiological sensing, and advanced ML architectures has created, for the first time, a technically feasible pathway toward objective, early, and clinically actionable chronic stress detection.

The BISP framework introduced here provides a structured clinical deployment architecture that addresses the full workflow from standardised specimen collection through laboratory quality control, ML feature engineering, model inference, and clinical output communication — converting laboratory ML capability into a implementable clinical diagnostic process. The framework's emphasis on confounder documentation, sex-stratified normalisation, phenotype classification alongside severity scoring, and SHAP-attributed individual explanations addresses the principal translational barriers that have limited clinical adoption of earlier biomarker-based stress assessments.

REFERENCES

[1] World Health Organization. (2022). World Mental Health Report: Transforming Mental Health for All. WHO, Geneva.

[2] Mohammadi, S. A. (2020). Integrative Approaches in the Management of Anxiety and Depression: Comparing Standard Pharmacotherapy with Combined Cognitive Behavioral Therapy and Adjunct Holistic Interventions. Journal of Applied Pharmaceutical Sciences and Research, 3(3), 21-33.

[3] McEwen, B. S. (1998). Stress, adaptation, and disease: Allostasis and allostatic load. Annals of the New York Academy of Sciences, 840(1), 33–44.

[4] Nalluri, S. K., Parasaram, V. K. B., & Bathini, V. T. (2021). Autonomous Manufacturing Operations Using Intelligent MES and Cloud-Native Analytics. Journal of Multidisciplinary Knowledge, 1(1), 45–55. Retrieved from https://jmk.datatablets.com/index.php/j/article/view/127

[5] Parasa, M. (2020). Control-mapped AI governance for high-risk HR decisions in SAP SuccessFactors: Audit-ready metrics for recruiting, performance calibration, and internal mobility. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 12(2), 153–168. https://doi.org/10.18090/samriddhi.v12i02.15

[6] MARASANI, Y. (2024). Enterprise Readiness for Generative AI: The Critical Role of Data Engineering. Frontiers in Computer Science and Artificial Intelligence, 3(2), 59-71.

[7] MARASANI, Y. (2023). Machine Learning Models for Predicting Patient Treatment Switching Using Claims Data. Frontiers in Computer Science and Artificial Intelligence, 2(1), 59-66.

[8] Mohammadi, S. A. (2023). The Role of Holistic Lifestyle Interventions (Mindfulness, Nutrition, Sleep Optimization and Traditional Therapies) in Improving Stress-Related Disorders and Quality of Life: A Systematic Review. INTERNATIONAL JOURNAL OF APPLIED PHARMACEUTICAL SCIENCES AND RESEARCH, 8(02), 42-54.

[9] Satish Kumar Nalluri, Venkata Krishna Bharadwaj Parasaram, Varun Teja Bathini. (2020). Secure Automation Frameworks for Smart Manufacturing Using Blockchain-Assisted Traceability. International Journal of Research & Technology, 8(2), 47–53. Retrieved from https://ijrt.org/j/article/view/879

[10] Mohammadi, S. A. (2022). Effects of Chronic Stress on Neuroendocrine and Immune Function: Clinical Implications for Early Intervention in Psychosomatic Disorders. Journal of Advanced Scientific Research, 13(03), 206-219.

[11] Satish Kumar Nalluri, Venkata Krishna Bharadwaj Parasaram. (2019). Software-Centric Automation Frameworks Integrating AI and Cybersecurity Principles. International Journal of Engineering Science & Humanities, 9(1), 30–40. Retrieved from https://www.ijesh.com/j/article/view/539

[12] Segerstrom, S. C., & Miller, G. E. (2004). Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry. Psychological Bulletin, 130(4), 601–630.

[13] Aardal, E., & Holm, A. C. (1995). Cortisol in saliva — Reference ranges and relation to cortisol in serum. European Journal of Clinical Chemistry and Clinical Biochemistry, 33(12), 927–932.

[14] Stalder, T., & Kirschbaum, C. (2012). Analysis of cortisol in hair: State of the art and future directions. Brain, Behavior, and Immunity, 26(7), 1019–1029.

[15] Parasa, M. (2022). Addressing the underutilization of exit interview data: A structured AI-assisted framework for actionable workforce insights in SAP SuccessFactors. Global Scientific and Academic Research Journal of Multidisciplinary Studies, 1(6), 42–52. https://gsarpublishers.com/abstract-2326/