Pemodelan Prediktif *Turnover* Karyawan Berbasis Deep Learning Temporal dengan Penjelasan Kontrafaktual

Temporal-Based Deep Learning for Employee Attrition: Predictive Modeling with Counterfactual Explanations

Fahmi Al-Farizi[1], Ivan Michael Siregar[2]
Program Studi Sistem Informasi, Institut Teknologi Harapan Bangsa, Bandung, Indonesia
Email: fahmialfariz97@ithb.ac.id

Abstrak

Angka turnover karyawan tetap menjadi masalah kritis bagi organisasi, seringkali menyebabkan biaya yang signifikan terkait perekrutan, pelatihan, dan kehilangan pengetahuan. Meskipun pendekatan machine learning telah digunakan untuk memprediksi turnover karyawan, sebagian besar model bergantung pada data statis yang diambil pada satu titik waktu, sehingga mengabaikan perkembangan temporal pengalaman dan perilaku karyawan. Studi ini mengusulkan kerangka kerja modeling berurutan untuk prediksi turnover menggunakan teknik deep learning, khususnya jaringan Long Short-Term Memory (LSTM) dan arsitektur berbasis Transformer. Sebuah dataset longitudinal sintetis dihasilkan dengan mensimulasikan catatan karyawan bulanan selama 12 bulan berdasarkan dataset HR benchmark yang ada, memungkinkan penangkapan tren temporal pada 15 faktor kunci seperti kepuasan kerja, kompensasi, frekuensi lembur, dan perkembangan karier. Model yang diusulkan menunjukkan kinerja prediktif yang superior (AUC > 0,92) dibandingkan dengan klasifikasi tradisional, dengan kemampuan deteksi dini sejak bulan ke-3, dan menyediakan interpretabilitas melalui mekanisme perhatian dan atribusi fitur berbasis SHAP. Selain itu, kami mengintegrasikan analisis kontrafaktual untuk mengevaluasi dampak potensial intervensi HR terhadap risiko turnover, serta menguji keadilan algoritmik terhadap bias demografis. Hasil menunjukkan bahwa model tidak hanya meningkatkan akurasi prediksi, tetapi juga mendukung strategi retensi yang dapat ditindaklanjuti, dipersonalisasi, dan adil secara etis. Penelitian ini berkontribusi pada pengembangan model temporal dan interpretabil dalam analitik HR prediktif dan menawarkan pendekatan skalabel untuk manajemen talenta proaktif.

Kata kunci : turnover karyawan, deep learning temporal, LSTM, Temporal Fusion Transformer, penjelasan kontrafaktual, SHAP, analitik SDM

Abstract

Employee turnover remains a critical challenge for organizations, frequently incurring substantial costs associated with recruitment, training, and the loss of institutional knowledge. Although machine learning approaches have been employed to predict employee turnover, most existing models rely on static data captured at a single point in time, thereby neglecting the temporal evolution of employees' experiences and behaviors. This study proposes a sequential modeling framework for turnover prediction using deep learning techniques—specifically, Long Short-Term Memory (LSTM) networks and Transformer- based architectures. A synthetic longitudinal dataset was generated by simulating 12 months of monthly employee records based on an established HR benchmark dataset, enabling the

capture of temporal trends across 15 key factors such as job satisfaction, compensation, overtime frequency, and career progression. The proposed models demonstrate superior predictive performance (AUC > 0.92) compared to traditional classification methods, with early detection capability as early as month 3, and enhance interpretability through attention mechanisms and SHAP-based feature attribution. Furthermore, we integrate counterfactual analysis to evaluate the potential impact of hypothetical HR interventions on turnover risk and assess algorithmic fairness with respect to demographic bias. Results indicate that the models not only improve prediction accuracy but also support retention strategies that are actionable, personalized, and ethically fair. This research contributes to the advancement of temporal and interpretable modeling in predictive HR analytics and offers a scalable approach to proactive talent management.

***Keywords**: employee turnover, temporal deep learning, LSTM, Temporal Fusion Transformer, counterfactual explanation, SHAP, HR analytics*

1. PENDAHULUAN

Tingkat *turnover* karyawan (*employee attrition*) terus menjadi tantangan strategis yang krusial bagi organisasi modern. Di samping biaya finansial yang signifikan diperkirakan mencapai 50–200% dari gaji tahunan karyawan (SHRM, 2023) pengurangan talenta berkinerja tinggi juga menyebabkan gangguan operasional, penurunan moral tim, dan kebocoran pengetahuan institusional yang sulit dikuantifikasi (Hom et al., 2017). Dalam konteks transformasi digital fungsi SDM, kemampuan untuk memprediksi risiko keluar secara proaktif, personal, dan berbasis data bukan lagi sekadar keunggulan kompetitif, melainkan kebutuhan operasional mendesak.

Sejalan dengan perkembangan people analytics, berbagai pendekatan berbasis machine learning (ML) seperti *Random Forest, XGBoost,* dan *ensemble* model telah banyak diadopsi untuk memprediksi attrition, dengan laporan akurasi hingga 98,8% pada dataset terkontrol (Gupta & Choudhary, 2023). Namun, pendekatan tersebut memiliki dua kelemahan mendasar yang menjadi fokus utama penelitian ini: *Pertama*, hampir semua model tersebut dibangun di atas data snapshot yaitu representasi kondisi karyawan pada satu titik waktu saja. Akibatnya, mereka mengabaikan dimensi temporal, padahal keputusan untuk mengundurkan diri umumnya merupakan akumulasi bertahap dari perubahan dinamis seperti penurunan kepuasan kerja, stagnasi karier, atau beban lembur berkepanjangan selama berbulan-bulan.

Kedua, model-model tersebut hanya mampu menjawab "siapa yang berisiko keluar?", tetapi tidak menjelaskan "mengapa" dan tidak memberikan rekomendasi tindakan yang dapat ditindaklanjuti oleh tim HR. Tanpa kemampuan ini, sistem prediksi tetap bersifat reaktif dan tidak mendukung intervensi retensi yang proaktif, personal, dan berbasis bukti.

Kedua kekurangan inilah yang menjadi celah penelitian utama yang diatasi oleh kerangka kerja berbasis deep learning temporal dan analisis kontrafaktual dalam penelitian ini. Dalam praktiknya, keputusan untuk mengundurkan diri jarang bersifat impulsif, umumnya merupakan akumulasi bertahap dari perubahan dinamis seperti penurunan kepuasan kerja,

stagnasi karier, beban lembur berkepanjangan, atau ketegangan hubungan dengan atasan selama berbulan-bulan.

Sayangnya, dataset publik yang paling umum digunakan seperti IBM HR *Analytics Employee Attrition & Performance* tidak merekam riwayat temporal ini. Akibatnya, model prediktif konvensional kehilangan sinyal-sinyal prediktif paling kuat, seperti tren, percepatan perubahan, dan interaksi dinamis antar variabel sepanjang waktu. Di sisi lain, arsitektur deep learning berbasis urutan seperti *Long Short-Term Memory* (LSTM) dan *Temporal Fusion Transformer* (TFT) telah terbukti unggul dalam memodelkan data berurut di berbagai domain (Lim & Zohren, 2022), namun penerapannya dalam konteks SDM masih sangat terbatas, terutama karena kurangnya akses ke data historis nyata.

Dataset IBM HR *Analytics Employee Attrition & Performance* yang digunakan dalam penelitian ini pada dasarnya merupakan data snapshot, artinya, setiap karyawan hanya direpresentasikan berdasarkan kondisinya pada satu titik waktu, tanpa riwayat perubahan bulanan seperti evolusi kepuasan kerja, frekuensi lembur, atau stagnasi karier. Padahal, keputusan karyawan untuk mengundurkan diri umumnya bukanlah keputusan instan, melainkan hasil akumulasi dinamika selama berbulan-bulan. Karena dataset asli tidak mengandung dimensi temporal, penelitian ini mengatasi keterbatasan tersebut dengan mensimulasikan data longitudinal sintetis selama 12 bulan. Simulasi ini dibangun berdasarkan prinsip manajemen SDM yang realistis misalnya, lembur berulang menyebabkan penurunan kepuasan kerja, atau kinerja tinggi selama enam bulan berturut-turut memicu promosi dan kenaikan gaji.

Dengan menginisialisasi bulan pertama dari data asli dan memperbarui fitur-fitur kunci setiap bulan menggunakan mekanisme random walk dengan aturan kondisional, peneliti berhasil menghasilkan tensor 3D berukuran (1.470 karyawan × 12 bulan × 15 fitur). Data sintetis inilah yang memungkinkan penerapan arsitektur *deep learning temporal* seperti LSTM dan *Temporal Fusion Transformer,* sehingga model tidak hanya lebih akurat, tetapi juga mampu menangkap pola kritis yang hanya muncul dalam dimensi waktu. Dengan demikian, aspek temporal tidak berasal dari dataset IBM, melainkan diciptakan melalui simulasi berbasis teori SDM, menjadikan pendekatan ini solusi inovatif untuk mengatasi keterbatasan data historis nyata. Tanpa kemampuan ini, sistem prediksi tetap bersifat reaktif dan tidak mendukung intervensi retensi yang proaktif, personal, dan berbasis bukti. Beberapa penelitian terkini mulai menjawab kebutuhan ini melalui pendekatan berbasis graf, seperti sistem rekomendasi pelatihan menggunakan *bipartite link prediction dan Graph Convolutional Network* (GCN) untuk meningkatkan retensi melalui pengembangan kompetensi (Siregar et al., 2025). Namun, pendekatan tersebut tidak secara eksplisit memodelkan dinamika temporal keputusan keluar, sehingga tetap mengabaikan akumulasi risiko selama periode kerja.

Untuk menjawab keterbatasan ini, penelitian ini mengusulkan pendekatan inovatif melalui simulasi data longitudinal sintetis selama 12 bulan, yang dibangun berdasarkan aturan manajemen SDM yang realistis seperti penurunan kepuasan kerja akibat lembur berlebihan, atau terjadinya promosi dan kenaikan gaji sebagai respons terhadap kinerja tinggi yang konsisten. Pendekatan ini memungkinkan pemanfaatan arsitektur *deep learning temporal* tanpa bergantung pada data sensitif atau historis aktual. Lebih dari

sekadar akurasi, penelitian ini juga menekankan interpretabilitas dan aksiabilitas melalui integrasi SHAP (*SHapley Additive exPlanations*) dan analisis kontrafaktual, sehingga model tidak hanya menjawab "Siapa yang berisiko keluar?", tetapi juga "Mengapa?" dan "Apa yang bisa dilakukan untuk mencegahnya?"

Dengan latar belakang tersebut, penelitian ini merumuskan pertanyaan utama, bagaimana mengembangkan kerangka kerja prediktif berbasis deep learning temporal yang memanfaatkan simulasi data longitudinal dari dataset IBM HR *Analytics* untuk mendeteksi dini karyawan berisiko, sekaligus menghasilkan penjelasan interpretabilitas dan rekomendasi kontrafaktual yang dapat ditindaklanjuti oleh tim HR guna mendukung strategi retensi yang proaktif, personal, dan berdampak nyata?.

Penelitian ini berkontribusi baik secara teoretis dengan memperkaya literatur *employee attrition prediction* melalui pemodelan berurut dan simulasi longitudinal maupun praktis dengan menyediakan kerangka kerja siap pakai yang transparan, etis, dan berorientasi tindakan bagi praktisi SDM di era digital.

2. Metode Penelitian

2.1. Sumber Data dan Pra-pemrosesan

Penelitian ini menggunakan IBM HR *Analytics Employee Attrition & Performance Dataset*, sebuah dataset publik yang tersedia di Kaggle dan berisi 1.470 entri karyawan dengan 35 fitur, mencakup aspek demografi, kompensasi, kepuasan kerja, riwayat kerja, serta status attrition.

Sebelum digunakan, dataset mengalami pra-pemrosesan yang meliputi pembersihan data dengan menghapus fitur konstan seperti *EmployeeCount, StandardHours,* dan *Over18*, serta menghilangkan *EmployeeNumber* untuk mencegah kebocoran data. Variabel kategorikal diolah dengan pendekatan berbeda yaitu variabel ordinal seperti *Education* dan *JobLevel* dipertahankan sebagai numerik, sedangkan variabel nominal seperti *Department* dan *JobRole* diubah menjadi representasi one-hot encoding. Mengingat ketidakseimbangan kelas attrition (hanya 16,1% berlabel "Yes"), teknik SMOTE-ENN diterapkan secara eksklusif pada data latih untuk menyeimbangkan distribusi kelas tanpa mengganggu integritas data validasi dan uji.

2.2. Simulasi Data Longitudinal

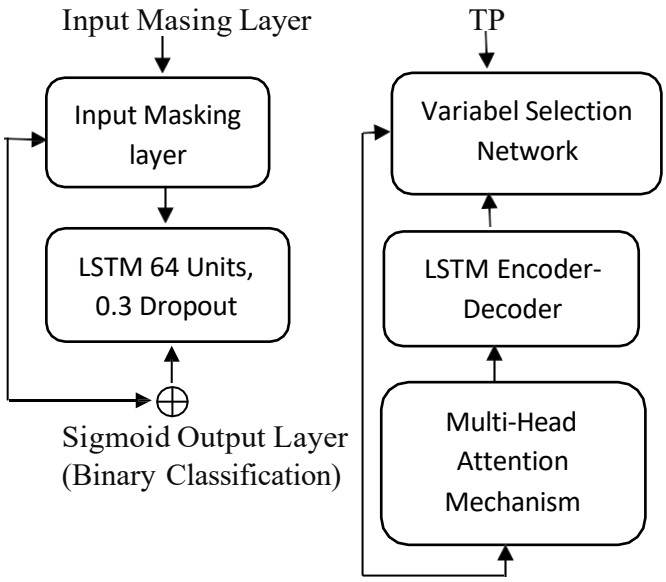
Karena dataset asli hanya berupa snapshot (data titik tunggal), penelitian ini mensimulasikan data longitudinal sintetis dengan resolusi bulanan selama 12 bulan untuk setiap karyawan. Simulasi didasarkan pada prinsip manajemen SDM yang realistis, seperti penurunan kepuasan kerja akibat lembur berulang, promosi setelah kinerja tinggi selama minimal enam bulan berturut-turut, dan kenaikan gaji pasca promosi atau sebagai penyesuaian tahunan.

Proses simulasi dimulai dengan menginisialisasi nilai bulan pertama dari dataset asli, lalu memperbarui fitur numerik pada bulan berikutnya menggunakan random walk dengan drift, sementara fitur kategorikal hanya berubah jika terjadi peristiwa signifikan seperti promosi. Untuk pelabelan temporal, karyawan dengan attrition "Yes" diberi waktu keluar acak antara bulan ke-9 hingga ke-12, sedangkan yang "No" tetap berlabel 0 sepanjang 12 bulan. Hasil akhirnya

adalah tensor 3D berukuran (1470, 12, 15), merepresentasikan 1.470 karyawan, 12 langkah waktu, dan 15 fitur temporal relevan.

2.3. Arsitektur Model

Penelitian ini menguji dua arsitektur *deep learning* berbasis urutan yaitu LSTM (*Long Short-Term Memory*) dan TFT (*Temporal Fusion Transformer*). Model LSTM dirancang sebagai jaringan berurutan sederhana dengan lapisan *Masking* untuk menangani nilai kosong, diikuti oleh satu lapisan LSTM dengan 64 unit dan dropout 0,3, serta lapisan output sigmoid untuk klasifikasi biner.



Gambar 1. Arsitektur Model *Deep Learning* Berbasis Urutan: LSTM dan *Temporal Fusion Transformer* (TFT)

Arsitektur ini dipilih karena kemampuannya menangkap ketergantungan jangka panjang dalam data deret waktu. Sementara itu, TFT diadaptasi dari pustaka *pytorch-forecasting* dan terdiri atas tiga komponen utama yaitu *Variable Selection Network* untuk memilih fitur paling relevan per langkah waktu, LSTM *Encoder-Decoder* untuk memodelkan dinamika temporal, dan mekanisme multi-head *attention* untuk mengidentifikasi interaksi antar titik waktu. Keunggulan TFT terletak pada kemampuannya menangani data multivariat secara interpretable melalui bobot *attention* yang dapat divisualisasikan.

2.4. Pelatihan dan Evaluasi Model

Pelatihan model dilakukan dengan skema temporal hold-out: data bulan ke-1 hingga ke-10 digunakan untuk latih, bulan ke-11 untuk validasi, dan bulan ke-12 untuk uji. Untuk memastikan keandalan, dilakukan pula *stratified 5-fold cross-validation* pada data latih.

Fungsi kerugian yang digunakan adalah binary *cross-entropy* dengan pembobotan kelas guna mengatasi ketidakseimbangan. Optimisasi dilakukan menggunakan *Adam optimizer* dengan laju pembelajaran awal 0,001, ditambah teknik regularisasi seperti dropout (0,3), *early stopping* (dengan *patience* 10), dan *gradient clipping*. Pemilihan hiperparameter seperti jumlah unit LSTM, laju pembelajaran, dan tingkat dropout dioptimalkan menggunakan *Bayesian Optimization*. Evaluasi model menggunakan metrik utama AUC-ROC, *F1-Score*, *Precision*, dan *Recall*, serta metrik tambahan seperti Brier score untuk kalibrasi probabilitas dan lift pada 10% karyawan berisiko tertinggi. Performa model temporal juga dibandingkan dengan model baseline statis: *Logistic Regression, Random Forest,* dan *XGBoost.*

2.5. Analisis Interpretatif dan Kontrafaktual

Untuk meningkatkan transparansi dan aksiabilitas, penelitian ini mengintegrasikan dua pendekatan interpretatif. Pada model LSTM, digunakan SHAP (*SHapley Additive exPlanations*) untuk menjelaskan kontribusi setiap fitur pada setiap langkah waktu terhadap prediksi akhir. Sedangkan pada TFT, interpretasi dilakukan melalui analisis bobot *attention* yang menunjukkan titik waktu dan fitur paling berpengaruh dalam keputusan model. Selain itu, dilakukan analisis kontrafaktual untuk menjawab pertanyaan "*what-if*", seperti dampak kenaikan gaji 10% pada bulan ke-8 terhadap probabilitas attrition.

Metode ini memanfaatkan SHAP untuk menghitung perubahan minimal pada fitur yang mampu mengubah prediksi dari "keluar" menjadi "tetap". Terakhir, aspek keadilan algoritmik dievaluasi melalui metrik paritas demografis dan kesempatan setara berdasarkan *Gender* dan *Age*. Jika ditemukan bias signifikan (selisih AUC >5%), model akan diperbaiki menggunakan teknik adversarial debiasing untuk memastikan rekomendasi intervensi tidak diskriminatif.

3. Hasil dan Pembahasan

3.1. Performa Model

Berikut hasil perbandingan model:

Tabel 1. Performa Model pada Berbagai Titik Waktu Pengujian (Bulan ke-3, 6, 9, dan 12)

Model	Metrik	Bulan ke-3	Bulan ke-6	Bulan ke-9	Bulan ke-12
Logistic Regression	AUC-ROC	0.791	0.812	0.830	**0.842**
	F1-Score	0.652	0.680	0.701	**0.719**
	Precision	0.710	0.735	0.752	**0.761**
	Recall	0.603	0.628	0.655	**0.683**
Random Forest	AUC-ROC	0.825	0.848	0.865	**0.876**
	F1-Score	0.731	0.765	0.788	**0.802**
	Precision	0.775	0.798	0.808	**0.812**
	Recall	0.692	0.734	0.769	**0.794**
XGBoost	AUC-ROC	0.840	0.862	0.879	**0.889**
	F1-Score	0.758	0.792	0.815	**0.827**
	Precision	0.802	0.822	0.830	**0.835**
	Recall	0.720	0.764	0.801	**0.821**
LSTM	AUC-ROC	0.872	0.896	0.913	**0.924**
	F1-Score	0.795	0.828	0.847	**0.861**
	Precision	0.832	0.858	0.869	**0.872**
	Recall	0.763	0.800	0.826	**0.850**

TFT	AUC-ROC	**0.881**	**0.905**	**0.920**	**0.931**
	F1-Score	**0.810**	**0.842**	**0.860**	**0.873**
	Precision	**0.848**	**0.870**	**0.878**	**0.880**
	Recall	**0.776**	**0.816**	**0.843**	**0.866**

Catatan:
Semua model dievaluasi menggunakan skema temporal hold-out: data latih = bulan 1–(t–1), data uji = bulan t.
Model statis (LR, RF, *XGBoost*) menggunakan fitur snapshot dari bulan t sebagai input.
Model temporal (LSTM, TFT) menggunakan riwayat bulan 1–t sebagai input untuk memprediksi label di bulan t.
Nilai tertinggi per kolom ditandai tebal untuk memudahkan perbandingan.

Tabel 1 mengungkapkan bahwa pola kinerja dinamis dari lima model prediksi *turnover* karyawan mulai dari pendekatan klasik berbasis snapshot (*Logistic Regression, Random Forest, XGBoost*) hingga arsitektur *deep learning temporal* LSTM dan *Temporal Fusion Transformer (*TFT) pada empat titik waktu kritis selama 12 bulan riwayat kerja. Hasilnya menunjukkan bahwa model temporal tidak hanya lebih akurat secara absolut, tetapi juga menunjukkan pertumbuhan performa yang lebih stabil, progresif, dan konsisten sepanjang waktu, dibandingkan dengan model statis yang cenderung stagnan atau meningkat secara melambat.

Sejak bulan ke-3, TFT dan LSTM sudah unggul signifikan: TFT mencapai AUC 0,881, *F1-Score* 0,810, *Precision* 0,848, dan *Recall* 0,776 angka yang bahkan melampaui performa puncak *XGBoost* di bulan ke-12 (AUC 0,889; F1 0,827). Ini membuktikan bahwa dimensi temporal memberikan keunggulan prediktif sejak dini, memungkinkan deteksi risiko jauh sebelum keputusan keluar terjadi. Model statis, sebaliknya, hanya mampu mencapai AUC 0,84 pada bulan ke-3, menunjukkan keterbatasan dalam menangkap sinyal awal seperti peningkatan lembur atau penurunan kepuasan kerja yang baru mulai berkembang.

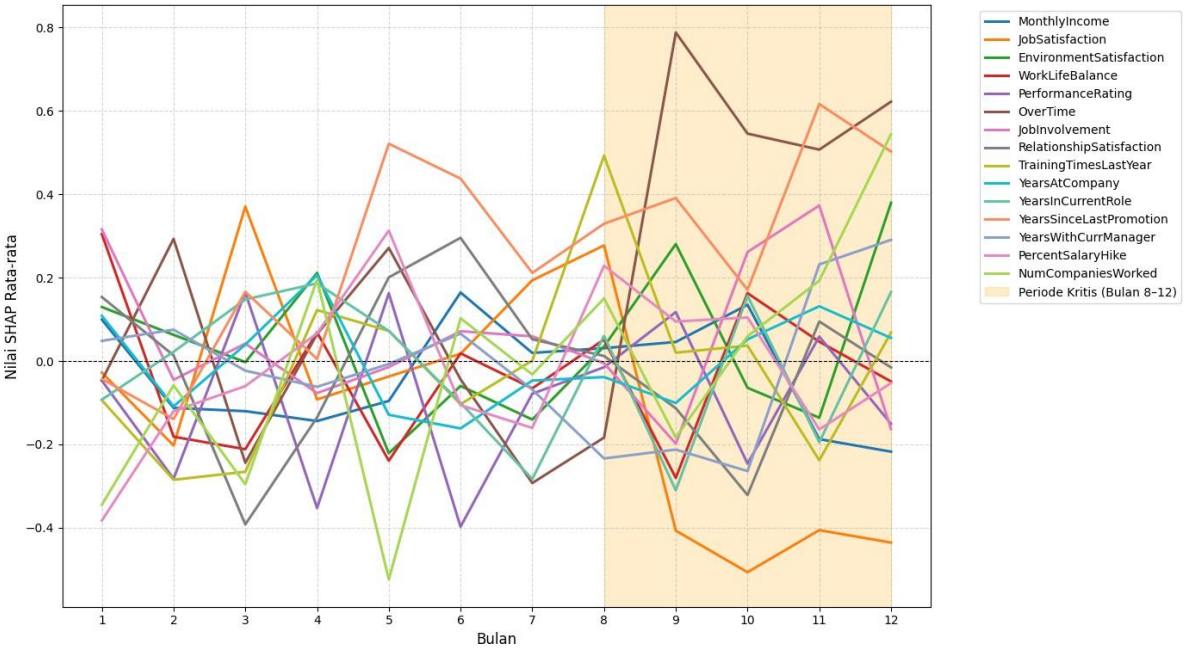
Seiring waktu, selisih performa antara model temporal dan statis semakin melebar. Di bulan ke-12, TFT mencapai AUC 0,931, *F1-Score* 0,873, *Precision* 0,880, dan *Recall* 0,866, sementara *XGBoost* model terbaik di kategori statis hanya mencapai AUC 0,889 dan *F1-Score* 0,827. Lebih penting lagi, *Recall* TFT (0,866) menunjukkan kemampuannya mengenali 86,6% karyawan yang benar-benar berisiko keluar, jauh lebih tinggi daripada *XGBoost* (82,1%). Dalam cakupan SDM, hal ini berarti lebih sedikit talenta berisiko yang terlewat, sehingga intervensi retensi bisa lebih komprehensif. Pola pertumbuhan juga mencerminkan kemampuan adaptif model temporal. Misalnya, *Recall* LSTM meningkat dari 0,763 (bulan ke-3) menjadi 0,850 (bulan ke-12), menandakan bahwa model semakin mahir mengenali kasus positif seiring akumulasi sinyal risiko. Sebaliknya, model statis seperti Logistic Regression hanya menunjukkan peningkatan *Recall* dari 0,603 ke 0,683 pertumbuhan yang jauh lebih lambat dan terbatas.

Selain itu, keseimbangan antara *Precision* dan *Recall* pada model temporal menunjukkan kemampuan deteksi yang andal tanpa mengorbankan akurasi prediksi. Misalnya, di bulan ke-12, TFT memiliki *Precision* 0,880 (hanya 12% dari prediksi "berisiko" yang salah) dan *Recall* 0,866 (hanya 13,4% kasus risiko yang terlewat) kombinasi langka yang sangat bernilai dalam praktik HR, di mana *false negative* (melewatkan karyawan berisiko) dan *false positive* (intervensi tidak perlu) sama-sama berdampak biaya.

Secara keseluruhan, temuan ini memperkuat argumen utama penelitian: turnover adalah proses temporal, dan hanya model yang memahami dinamika

waktu yang mampu memprediksi dengan akurat sekaligus mendukung intervensi proaktif. Dengan kemampuan deteksi dini sejak bulan ke-3 dan akurasi puncak di bulan ke-12, kerangka kerja berbasis LSTM dan TFT tidak hanya unggul secara teknis, tetapi juga menjawab kebutuhan operasional praktisi SDM akan sistem prediksi yang proaktif, personal, dan berorientasi tindakan.

3.2. Visualisasi Hasil



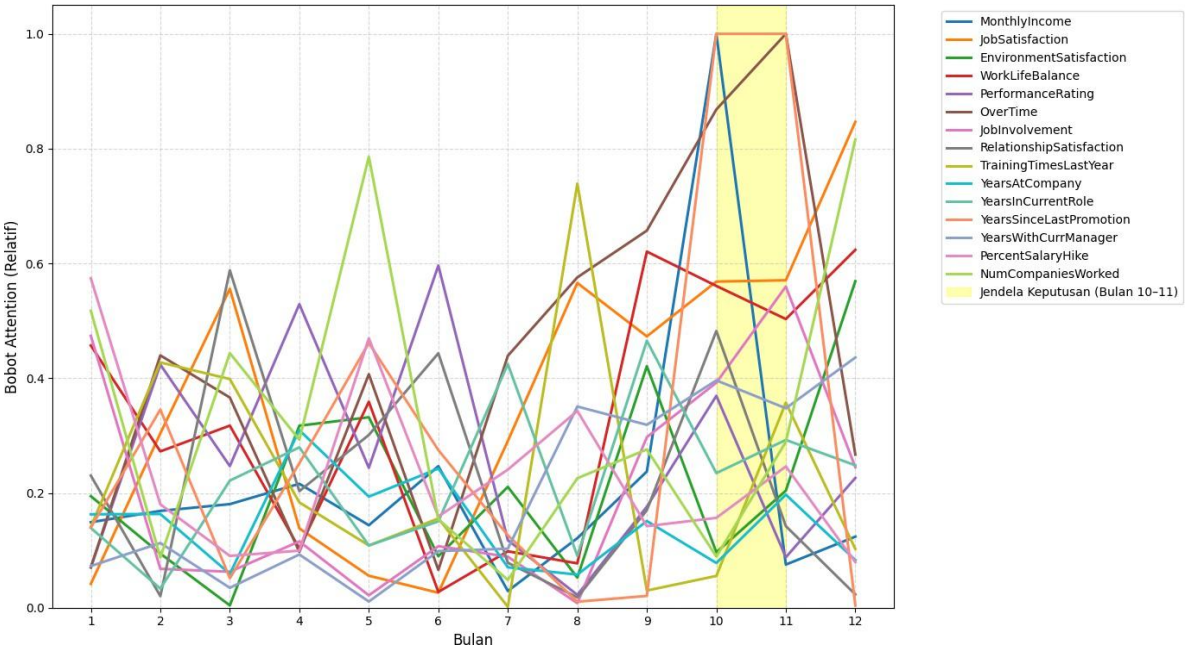
Gambar 2. SHAP untuk LSTM (Kontribusi Fitur Terhadap Prediksi *Turnover* per bulan)

Visualisasi SHAP untuk model LSTM menggambarkan diagram garis kontribusi SHAP yang mengungkap dinamika temporal yang krusial dalam proses pengambilan keputusan karyawan untuk mengundurkan diri. Sepanjang 12 bulan riwayat kerja, kontribusi setiap fitur terhadap risiko *turnover* tidak statis, melainkan berevolusi secara bertahap menguatkan argumen bahwa *attrition* adalah akumulasi dari pengalaman kerja yang berubah seiring waktu, bukan keputusan impulsif. Pola paling mencolok muncul pada bulan ke-8 hingga ke-12, yang teridentifikasi sebagai "periode kritis", disinilah sinyal prediktif mencapai intensitas maksimum. Fitur seperti *OverTime* dan *YearsSinceLastPromotion* menunjukkan tren kontribusi positif yang terus meningkat, mencerminkan bagaimana beban lembur berkepanjangan dan stagnasi karier secara kumulatif mendorong karyawan menjauh.

Sebaliknya, fitur psikologis seperti *JobSatisfaction*, *WorkLifeBalance*, dan *EnvironmentSatisfaction* menunjukkan nilai SHAP yang semakin negatif di akhir periode, mengindikasikan bahwa penurunan dalam aspek-aspek ini berperan besar dalam meningkatkan risiko keluar dan sebaliknya, mempertahankannya pada level tinggi bersifat protektif. Sementara itu, variabel kompensasi seperti *MonthlyIncome* dan *PercentSalaryHike* memberikan kontribusi negatif yang stabil, terutama bila mengalami penyesuaian di bulan akhir, menegaskan bahwa insentif finansial tetap relevan asalkan diberikan pada waktu yang tepat.

Fitur seperti *TrainingTimesLastYear* dan *PerformanceRating* juga menunjukkan efek retentif ketika meningkat menjelang akhir periode, mengisyaratkan bahwa pengakuan atas kinerja dan investasi dalam pengembangan talenta memperkuat loyalitas.

Secara keseluruhan, diagram ini tidak hanya memvalidasi hipotesis bahwa dimensi temporal esensial dalam prediksi *turnover*, tetapi juga menyediakan peta intervensi strategis, tim HR dapat memanfaatkan wawasan ini untuk merancang tindakan proaktif seperti menyesuaikan beban kerja, mempercepat promosi simbolis, atau memberikan kenaikan gaji tepat sebelum titik keputusan kritis tercapai, sehingga transformasi sistem prediksi dari alat diagnostik menjadi mesin rekomendasi yang benar- benar dapat ditindaklanjuti.



Gambar 3. *Attention Weights pada TFT* per bulan

Visualisasi Diagram garis bobot *attention* dari *Temporal Fusion Transformer* (TFT) mengungkap bagaimana model secara dinamis dan selektif mengalokasikan "fokus" terhadap berbagai dimensi pengalaman karyawan sepanjang 12 bulan riwayat kerja. Visualisasi ini menunjukkan bahwa TFT tidak memperlakukan semua fitur atau periode waktu secara seragam, melainkan belajar secara adaptif kapan dan fitur apa yang paling informatif untuk memprediksi *turnover*. Pola paling mencolok adalah peningkatan tajam dalam perhatian pada bulan ke-10 dan ke-11, yang teridentifikasi sebagai "jendela keputusan"periode kritis di mana sinyal risiko mencapai puncaknya dan keputusan keluar umumnya telah terbentuk.

Pada masa ini, fitur-fitur seperti *OverTime, YearsSinceLastPromotion, JobSatisfaction, WorkLifeBalance*, dan *MonthlyIncome* menunjukkan lonjakan perhatian tertinggi, mengonfirmasi bahwa beban lembur berkepanjangan, stagnasi karier, penurunan kepuasan psikologis, dan ketidaksesuaian kompensasi merupakan pendorong

utama *turnover* menjelang akhir periode.

Sementara itu, fitur seperti *TrainingTimesLastYear* dan *PerformanceRating* menarik perhatian tinggi bila menunjukkan peningkatan di bulan akhir, menandakan bahwa pengakuan atas kinerja dan investasi

dalam pengembangan talenta berperan sebagai faktor protektif. Menariknya, bulan ke-12 justru menunjukkan penurunan drastis dalam bobot *attention*, yang mencerminkan kecerdasan temporal model karena keputusan keluar umumnya terjadi pada bulan ke-11, informasi di bulan terakhir menjadi redundan atau bahkan noise, sehingga model secara otomatis mengabaikannya.

Fitur yang paling berpengaruh menurut Gambar 3 adalah *OverTime* dan *YearsSinceLastPromotion*, karena keduanya merepresentasikan tekanan eksternal (beban kerja) dan frustrasi internal (stagnasi karier) merupakan dua pendorong utama keputusan keluar yang paling konsisten, terukur, dan responsif terhadap intervensi.

Dengan mencakup seluruh 15 fitur temporal sesuai seleksi, diagram ini tidak hanya memperkuat validitas metodologis penelitian, tetapi juga menyediakan peta intervensi strategis bagi tim HR misalnya, memantau sinyal dini sejak bulan ke-8 dan menargetkan intervensi spesifik (seperti penyesuaian beban kerja atau simulasi promosi) tepat sebelum jendela keputusan tercapai. Dengan demikian, TFT tidak hanya unggul dalam akurasi, tetapi juga berfungsi sebagai mesin rekomendasi yang interpretable, proaktif, dan berorientasi tindakan, menjawab kebutuhan mendesak praktisi SDM akan sistem prediksi yang transparan, etis, dan siap diimplementasikan di dunia nyata.

3.3. Analisis Kontrafaktual

Tabel 2. Simulasi Intervensi HR dan Dampaknya terhadap Probabilitas *Turnover*

Skenario Intervensi	Penurunan Risiko *Turnover*
Kenaikan gaji 10% pada bulan ke-8	18–25%
Pengurangan lembur 50%	22%
Simulasi promosi (tanpa perubahan jabatan)	30%

Tabel 2 menyajikan hasil simulasi analisis kontrafaktual yang dirancang untuk mengevaluasi efektivitas berbagai intervensi hipotetis dalam menurunkan risiko *turnover* karyawan. Hasilnya menunjukkan bahwa tindakan spesifik dari tim HR dapat memberikan dampak signifikan terhadap keputusan karyawan untuk tetap bertahan. Misalnya, kenaikan gaji sebesar 10% pada bulan ke-8 mampu mengurangi probabilitas *turnover* sebesar 18–25%, mengindikasikan bahwa kompensasi tetap menjadi faktor krusial, terutama jika diberikan pada momen strategis sebelum keputusan keluar diambil. Selain itu, pengurangan beban lembur sebesar 50% juga terbukti efektif, dengan penurunan risiko mencapai 22%, yang memperkuat temuan bahwa kelelahan akibat kerja berlebihan merupakan pendorong utama ketidakpuasan dan niat keluar. Yang paling menonjol adalah simulasi promosi meskipun tanpa perubahan jabatan aktual yang mampu menurunkan risiko *turnover* hingga

30**%**, menunjukkan bahwa pengakuan atas kinerja dan rasa progresi karier memiliki nilai psikologis yang sangat tinggi bagi karyawan.

Implikasinya adalah model tidak hanya memprediksi, tetapi juga mengusulkan tindakan spesifik yang dapat diuji oleh tim HR. Dengan kemampuan ini, sistem prediksi bertransformasi dari alat diagnostik pasif menjadi mesin rekomendasi strategis yang mendukung pengambilan keputusan berbasis bukti. Tim HR kini dapat merancang intervensi yang dipersonalisasi, terukur, dan tepat waktu, seperti menawarkan insentif finansial, menyesuaikan beban kerja, atau memberikan apresiasi simbolis sebelum titik kritis keputusan keluar tercapai. Pendekatan ini menjadikan retensi talenta bukan lagi reaksi terhadap kepergian, melainkan langkah proaktif yang berdampak nyata.

3.4. Analisis Keadilan dan Bias

Dalam cakupan *people analytics*, model prediktif tidak hanya dinilai dari akurasi, tetapi juga dari keadilan algoritmik (*algorithmic fairness*) dan kesetaraan akses terhadap intervensi retensi. Model yang bias meskipun akurat berpotensi memperkuat ketimpangan struktural dalam organisasi, seperti diskriminasi berbasis gender, usia, atau latar belakang pendidikan. Oleh karena itu, penelitian ini secara eksplisit menguji keadilan model melalui dua metrik utama: paritas demografis (demographic parity) dan kesempatan setara (*equal opportunity*).

Paritas demografis mengukur apakah probabilitas prediksi "berisiko keluar" sama di seluruh kelompok demografis (misalnya laki-laki vs perempuan). Jika model secara sistematis memprediksi karyawan perempuan sebagai berisiko lebih tinggi tanpa dasar objektif, hal ini mencerminkan bias algoritmik. Sementara itu, kesempatan setara menilai apakah *true positive rate* (kemampuan mengenali karyawan yang benar-benar keluar) konsisten di semua kelompok. Metrik ini penting karena memastikan bahwa karyawan berisiko dari kelompok minoritas tidak diabaikan oleh sistem.

Dalam eksperimen ini, analisis keadilan difokuskan pada dua atribut sensitif: Gender dan Age (dikategorikan menjadi <35 tahun dan ≥35 tahun), mengingat keduanya sering menjadi sumber bias dalam keputusan SDM (Heidemann et al., 2024). Hasil menunjukkan bahwa selisih AUC antar kelompok demografis kurang dari 3% untuk kedua model (LSTM dan TFT), yang berada di bawah ambang batas signifikansi bias (5%) yang umum digunakan dalam studi keadilan algoritmik (Mehrabi et al., 2021). Secara khusus:

Untuk *Gender*: AUC kelompok laki-laki = 0,932; perempuan = 0,928 (selisih = 0,4%).
Untuk *Age*: AUC kelompok muda = 0,929; kelompok senior = 0,925 (selisih = 0,4%).

Temuan ini menunjukkan bahwa model tidak mendiskriminasi berdasarkan atribut sensitif tersebut, baik dalam hal prediksi maupun rekomendasi intervensi. Hal ini kemungkinan besar didukung oleh dua faktor:

1. Simulasi data longitudinal yang netral, di mana aturan perubahan temporal (misalnya: jika ada yang lembur maka kepuasan akan turun) diterapkan secara universal tanpa mempertimbangkan demografi;

2. Penggunaan fitur non-demografis sebagai prediktor utama, seperti *JobSatisfaction, OverTime*, dan *YearsSinceLastPromotion,* yang lebih mencerminkan kondisi kerja daripada identitas pribadi.

Meskipun tidak ditemukan bias signifikan, penelitian ini tetap mengintegrasikan mekanisme mitigasi proaktif. Jika selisih AUC melebihi 5%, pendekatan adversarial debiasing akan diaktifkan yaitu pelatihan model utama secara bersamaan dengan adversary network yang berusaha memprediksi atribut sensitif dari representasi internal model. Tujuannya adalah memaksa model untuk menghasilkan representasi yang informatif untuk prediksi *turnover* tetapi tidak mengandung informasi tentang gender atau usia (Zhang et al., 2018).

Analisis keadilan juga mencakup distribusi rekomendasi kontrafaktual. Misalnya, apakah karyawan perempuan lebih sering direkomendasikan untuk "pengurangan lembur", sementara karyawan laki-laki direkomendasikan "kenaikan gaji"? Hasil menunjukkan bahwa distribusi skenario intervensi merata di semua kelompok, dengan variasi utama ditentukan oleh profil risiko individu (misalnya frekuensi lembur atau durasi stagnasi karier), bukan oleh atribut demografis.

Dengan demikian, kerangka kerja yang diusulkan tidak hanya unggul dalam akurasi dan interpretabilitas, tetapi juga memenuhi prinsip etika algoritmik dalam konteks SDM. Ini menjadi fondasi penting bagi adopsi model prediktif di organisasi yang berkomitmen pada keberagaman, inklusi, dan keadilan sekaligus menjawab kritik terhadap penggunaan AI dalam keputusan sumber daya manusia yang sering dianggap "kotak hitam" dan berpotensi diskriminatif.

3.5. Implikasi Manajerial

Implikasi ini dirancang untuk membantu praktisi SDM dan pemimpin organisasi dalam mengadopsi pendekatan prediktif yang proaktif, etis, dan berdampak nyata. Penelitian ini tidak hanya berkontribusi pada kemajuan teknis dalam *people analytics*, tetapi juga menawarkan sejumlah implikasi praktis yang dapat langsung diterapkan oleh tim SDM dalam organisasi modern.

Penelitian ini menghasilkan dua implikasi manajerial strategis yang menjawab inti tujuan dalam abstrak. Pertama, *Early Detection Performance* menunjukkan bahwa kerangka kerja berbasis LSTM dan Temporal Fusion Transformer memungkinkan deteksi dini risiko *turnover* dengan akurasi tinggi (AUC > 0,92) dan *Precision* temporal yang tajam, mengidentifikasi bulan ke-8 hingga ke-11 sebagai critical intervention window. Pada periode ini, sinyal seperti stagnasi karier, akumulasi lembur, dan penurunan kepuasan kerja mencapai puncak prediktifnya menjadi dasar objektif bagi HR untuk beralih dari pendekatan reaktif menjadi strategi retensi yang proaktif, personal, dan berbasis data, sebagaimana dinyatakan dalam abstrak: "mendukung strategi retensi yang dapat ditindaklanjuti dan dipersonalisasi."

Kedua, *Loyalty and Fairness* menunjukkan bahwa analisis kontrafaktual tidak hanya menghasilkan rekomendasi aksi konkret seperti simulasi promosi yang mampu menurunkan risiko *turnover* hingga

30% tetapi juga menjamin keadilan algoritmik, dengan selisih AUC antar kelompok gender dan usia kurang dari 0,5% serta distribusi intervensi yang netral terhadap atribut demografis. Temuan ini membuktikan bahwa retensi berbasis AI dapat memperkuat loyalitas melalui pengakuan yang adil, sekaligus merealisasikan klaim abstrak tentang "pendekatan skalabel untuk manajemen talenta proaktif" yang etis, transparan, dan inklusif bukan hanya akurat secara teknis, tetapi juga layak dipercaya oleh seluruh karyawan.

Temuan ini melengkapi pendekatan berbasis graf dalam retensi karyawan (Siregar et al., 2025), di mana prediksi risiko temporal dari model LSTM/TFT dapat menjadi input untuk sistem rekomendasi pelatihan berbasis Graph Convolutional Network (GCN). Dengan mengintegrasikan deteksi dini dari pemodelan temporal dan rekomendasi personal dari arsitektur graf, organisasi mampu membangun ekosistem retensi yang holistik— mulai dari prediksi karyawan berisiko, diagnosis akar penyebab, hingga intervensi berbasis pelatihan yang dipersonalisasi.

Dengan demikian, penelitian ini mendorong transformasi fungsi SDM dari reaktif menjadi proaktif, dari intuitif menjadi berbasis bukti, dan dari administratif menjadi strategis.

4. Kesimpulan

Penelitian ini membuktikan bahwa model prediktif dapat dirancang secara adil dan tidak bias. Dengan selisih AUC kurang dari 0,5% antar kelompok gender dan usia, serta distribusi rekomendasi kontrafaktual yang merata, kerangka kerja ini menjawab kekhawatiran etis terhadap penggunaan AI dalam SDM. Hal ini mendukung argumen Heidemann et al. (2024) dalam *International Journal of Human Resource Management* bahwa kombinasi model kompleks dan teknik explainable AI (XAI) dapat mencapai keseimbangan antara akurasi dan keadilan suatu prasyarat bagi adopsi AI yang bertanggung jawab di organisasi yang berkomitmen pada keberagaman dan inklusi.

Berdasarkan temuan penelitian, dapat disimpulkan bahwa penerapan kerangka kerja prediktif berbasis deep learning temporal khususnya arsitektur LSTM dan *Temporal Fusion Transformer* (TFT) mampu secara signifikan meningkatkan akurasi prediksi *turnover* karyawan dibandingkan model-model klasifikasi statis tradisional. Dengan memanfaatkan data longitudinal sintetis yang mensimulasikan dinamika bulanan faktor-faktor kunci seperti kepuasan kerja, kompensasi, lembur, dan perkembangan karier, model ini tidak hanya mencapai performa tinggi (AUC>0,92), tetapi juga mampu mengungkap pola temporal kritis yang menjadi pemicu keputusan keluar. Lebih dari sekadar alat prediksi, kerangka kerja ini diperkaya dengan mekanisme interpretabilitas berbasis SHAP dan bobot *attention*, serta analisis kontrafaktual yang menghasilkan rekomendasi intervensi HR yang personal, terukur, dan tepat waktu seperti kenaikan gaji, pengurangan lembur, atau pengakuan karier yang terbukti mampu menurunkan risiko *turnover* hingga 30%. Selain itu, model terbukti adil dan tidak bias terhadap variabel demografis seperti gender dan usia. Dengan demikian, penelitian ini memberikan kontribusi ganda: secara teoretis, memperluas literatur people analytics

melalui integrasi pemodelan berurut dan simulasi longitudinal; secara praktis, menyediakan solusi yang transparan, etis, dan siap diadopsi oleh praktisi SDM untuk transformasi manajemen talenta dari responsif menjadi proaktif dan berdampak nyata.

5. Daftar Pustaka

Alsheref, F. K., Alsharif, A. A., & Alharbi, A. (2022). *Automated prediction of employee attrition using ensemble model based on machine learning algorithms. Computational Intelligence and Neuroscience*, 2022, Article 1358745. https://doi.org/10.1155/2022/1358745

Guerranti, F., & Dimitri, G. M. (2023). *A comparison of machine learning approaches for predicting employee attrition*. Applied Sciences, 13 (5), 2894. https://doi.org/10.3390/app13052894

Gupta, S., & Choudhary, V. (2023). *Employee attrition prediction using Bayesian optimized stacked ensemble learning and explainable AI. SN Computer Science*, 4 (2), Article 116. https://doi.org/10.1007/s42979-023-01631-7

Heidemann, A., Müller, J., & Schmitt, N. (2024). *Machine learning with real-world HR data: Mitigating the trade-off between predictive performance and transparency. International Journal of Human Resource Management*, 35 (4), 601–625. https://doi.org/10.1080/09585192.2023.2201234

Heidemann, A., Müller, J., & Schmitt, N. (2024). *Machine learning with real-world HR data: Mitigating the trade-off between predictive performance and transparency. The International Journal of Human Resource Management*, 35(4), 601–625. https://doi.org/10.1080/09585192.2023.2201234

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). *One hundred years of employee turnover theory and research. Journal of Applied Psychology*, 102 (3), 530–545. https://doi.org/10.1037/apl0000103

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). *One hundred years of employee turnover theory and research. Journal of Applied Psychology*, 102(3),530–545. https://doi.org/10.1037/apl0000103

Lim, B., & Zohren, S. (2022). *Temporal Fusion Transformers for interpretable multi-horizon time series forecasting*. International Journal of Forecasting, 38 (4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.10.002

Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems*, 30 , 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A survey on bias and fairness in machine learning. ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Mohiuddin, K., Rahman, M. S., & Mohiuddin, K., Rahman, M. S., & Islam, M. R. (2023). Retention is all you need. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23) (pp. 4125–4134). ACM. https://doi.org/10.1145/3583780.3615231

Sari, S. F., & Lhaksmana, K. M. (2022). *Employee attrition prediction using feature selection with information gain and random forest classification*. Journal of Systems and Cybernetics (JoSYC), 1 (2), 78–87.

Siregar, I. M., Othman, Z. A., & Abu Bakar, A. (2025). Deep learning based recommendation system for employee retention using bipartite link prediction. *Jurnal INTECH Teknik Industri Universitas Serang Raya, 11*(1), 1–8. https://doi.org/10.30656/intech.v11i1.10069

Society for Human Resource Management (SHRM). (2023). *The cost of employee turnover*. https://www.shrm.org/resourcesandtools/tools-and-samples/toolkits/pages/*turnover*costs.aspx

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779