

Popularitas dan Tren Metode Pemodelan *Expected Goals* (xG): Sebuah Analisis Bibliometrik

Fadhil Raihan Akbar^{*1}, Qurrotul Aini², Nida'ul Hasanati³

^{1,2,3}Universitas Islam Negeri Syarif Hidayatullah Jakarta; Jl. Ir H. Juanda No.95, Ciputat, Kota Tangerang Selatan, Banten 15412

Jurusan Sistem Informasi, Fakultas Sains dan Teknologi

e-mail: ^{*1}fadhil.raihanakbar21@mhs.uinjkt.ac.id, ²qurrotul.aini@uinjkt.ac.id,
³nidaul.hasanati@uinjkt.ac.id

Abstrak

Analisis data telah secara fundamental mengubah evaluasi kinerja dalam sepak bola kontemporer, dengan Expected Goals (xG) muncul sebagai salah satu metrik paling signifikan. Terlepas dari pemanfaatannya yang luas, dokumentasi menyeluruh mengenai perkembangan metodologi pemodelannya tetap langka. Upaya penelitian ini bercita-cita untuk menjelaskan lanskap ilmiah global yang berkaitan dengan metodologi pemodelan xG melalui penerapan kerangka bibliometrik. Data diekstraksi dari database Scopus dan dianalisis menggunakan VOSViewer dan perangkat lunak analitis tambahan untuk membedakan tren, metodologi umum, dan domain implementasi. Temuan ini menunjukkan peningkatan penting dalam volume publikasi dari tahun 2020 hingga 2025, menunjukkan minat ilmiah yang terus meningkat. Random Forest telah menjadi terkenal sebagai metodologi pemodelan yang paling banyak digunakan, meskipun tren saat ini mengungkapkan kebangkitan yang mendukung Regresi Logistik. Domain utama implementasi untuk Tujuan yang diharapkan (xG) berada di ranah Analisis Kinerja, kemudian digantikan oleh Analisis Taktis dan Strategis. Pelaksanaan analisis jaringan kata kunci telah menggambarkan tiga cluster dominan yang berfokus pada metodologi yang berkaitan dengan pembelajaran mesin, model regresi, dan teknik pembelajaran mendalam. Temuan penelitian ini menyajikan gambaran lengkap mengenai kondisi terkini dan arah masa depan riset xG, yang diharapkan dapat bermanfaat bagi para peneliti dan praktisi di bidang analisis olahraga.

Kata kunci— *Expected Goals*, xG, Analisis Bibliometrik, VOSviewer, Analisis Sepak Bola

Abstract

Data analysis has fundamentally altered the landscape of performance assessment in contemporary football, with Expected Goals (xG) emerging as one of the preeminent metrics. Notwithstanding its extensive utilization, a thorough examination of the progression of its modeling methodologies remains deficient. This investigation seeks to delineate the global research terrain concerning xG modeling methodologies through a bibliometric lens. Empirical data was extracted from the Scopus database and subsequently analyzed utilizing VOSviewer alongside other analytical tools to discern trends, prevalent methodologies, and areas of application. The findings indicate a pronounced escalation in the volume of publications between 2020 and 2025, signifying an increasing scholarly engagement. Random Forest is identified as the most prevalently employed modeling technique, although recent patterns suggest a revival in the favor of Logistic Regression. The principal domain of application for xG is Performance Analysis, succeeded by Tactical & Strategic Analysis. Keyword network analysis has unveiled three primary clusters focused on machine learning methodologies, regression techniques, and deep learning strategies. These results offer a lucid framework of the present condition and prospective trajectory of xG modeling research, furnishing significant insights for both scholars and practitioners within the realm of sports analytics.

Keywords— *Expected Goals*, xG, *Bibliometric Analysis*, VOSviewer, *Football Analytics*



This is an open-access article under the [CC-BY-CA](#) license

1. PENDAHULUAN

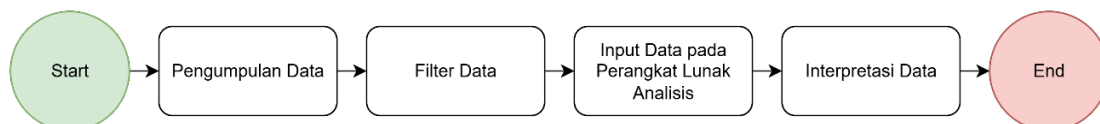
Banyak industri telah sangat terpengaruh oleh munculnya revolusi data, dan olahraga tidak terkecuali. Selama bertahun-tahun, analisis kinerja sepak bola telah berkembang dari metrik analitis yang lebih rumit yang memungkinkan pemahaman yang lebih mendalam dan prediktif menjadi statistik deskriptif konvensional, seperti menghitung gol dan upaya [1]. *Expected Goals* (xG) adalah salah satu metrik yang paling transformatif dalam beberapa tahun terakhir. Berdasarkan data historis dari ribuan tembakan serupa, metrik ini menentukan kualitas peluang tembakan dengan memberikan probabilitas bahwa tembakan tersebut akan menjadi gol [2]. Dalam perhitungannya, hal-hal seperti lokasi tembakan, sudut, jarak, dan jenis umpan sangat diperhatikan.

xG memiliki kemampuan untuk membedakan kualitas peluang dari hasil akhir (*finishing*). Hal ini memungkinkan untuk melakukan analisis yang lebih objektif terhadap kinerja serangan dan pertahanan seseorang atau tim, terlepas dari faktor keberuntungan atau perubahan singkat [3]. Metode pemodelan xG juga berkembang seiring dengan ketersediaan data granular yang lebih besar. Pendekatan statistik sederhana seperti regresi logistik sering digunakan dalam model awal [4], namun kini penelitian telah berkembang ke arah penggunaan algoritma *machine learning* yang lebih kompleks seperti *Random Forest*, *Gradient Boosting*, dan bahkan Jaringan Saraf Tiruan (*Neural Networks*) untuk menangkap interaksi non-linear antar variabel [5], [7], [8].

Meskipun telah banyak penelitian yang mengembangkan atau membandingkan model xG, masih terdapat celah dalam literatur yang secara komprehensif memetakan lanskap penelitian ini secara keseluruhan. Analisis bibliometrik menawarkan metode kuantitatif untuk mengevaluasi perkembangan suatu bidang ilmiah dengan menganalisis metadata publikasi [6]. Dengan menggunakan metode analisis bibliometrik, studi ini menyelidiki perkembangan penelitian terkait pemodelan xG. Tujuannya adalah untuk mengetahui bagaimana gaya riset xG, popularitas berbagai metode, serta area implementasinya telah berubah, sekaligus mengidentifikasi tren publikasi, kolaborasi, dan metode penelitian yang akan datang.

2. METODE PENELITIAN

Penelitian ini menggunakan analisis bibliometrik untuk melihat penelitian sebelumnya tentang pemodelan *Expected Goals* (xG). Proses pengumpulan data, filter data, analisis data, dan interpretasi hasil adalah langkah-langkah utama yang dilakukan dalam studi ini. Gambar 1 menunjukkan desain penelitian.



Gambar 1 Alur Penelitian

2.1 Pengumpulan Data

Dalam penelitian ini, pengumpulan data dilakukan dengan memanfaatkan *metadata* literatur dari basis data Scopus. Proses pencarian dilakukan menggunakan *string* kueri pencarian lanjutan yang komprehensif untuk memastikan cakupan hasil yang luas dan relevan. Kueri ini dirancang untuk menangkap publikasi yang mengandung istilah inti ("*expected goals*" atau "xg"),

dalam konteks domain spesifik ("*football*" atau "*soccer*"), dan secara eksplisit membahas salah satu dari serangkaian luas metode pemodelan statistik maupun *machine learning*.

2. 2 Filter Data

Setelah hasil pencarian awal dikumpulkan, proses penyaringan dilakukan untuk memilih produk yang memenuhi kriteria *Quality Assessment* (QA). QA digunakan untuk memastikan kualitas dan relevansi jurnal ilmiah yang dipilih sesuai dengan tujuan penelitian. Kriteria QA pada penelitian ini dapat dilihat pada Tabel 1.

Tabel 1 Kriteria *Quality Assessment*

QA	Kriteria
1	Penelitian dipublikasikan pada rentang waktu 2020-2025.
2	Jenis dokumen adalah <i>Article</i> atau <i>Conference Paper</i> .
3	Penelitian berfokus pada pengembangan atau implementasi model xG.

Data yang memenuhi kriteria Tabel 1 disaring menggunakan fitur filter yang tersedia di platform Scopus. Setelah itu, data yang lolos penyaringan diekstraksi dalam format BibTeX dan CSV untuk diproses oleh perangkat lunak analisis bibliometrik.

2. 3 Analisis Data

Pada penelitian ini, analisis data dilakukan dengan pendekatan bibliometrik untuk mengeksplorasi dan memvisualisasikan tren penelitian. Untuk menganalisis data bibliometrik, penelitian ini menggunakan perangkat lunak VOSviewer (versi 1.6.18). VOSviewer dipakai untuk membangun sebuah peta visual yang menunjukkan jaringan *co-occurrence* atau kemunculan bersama antar kata kunci. Melalui peta visual tersebut, kluster-kluster topik utama dalam literatur bisa diidentifikasi dan hubungan antar konsep dapat dipetakan dengan lebih jelas.

2.4. Interpretasi Data

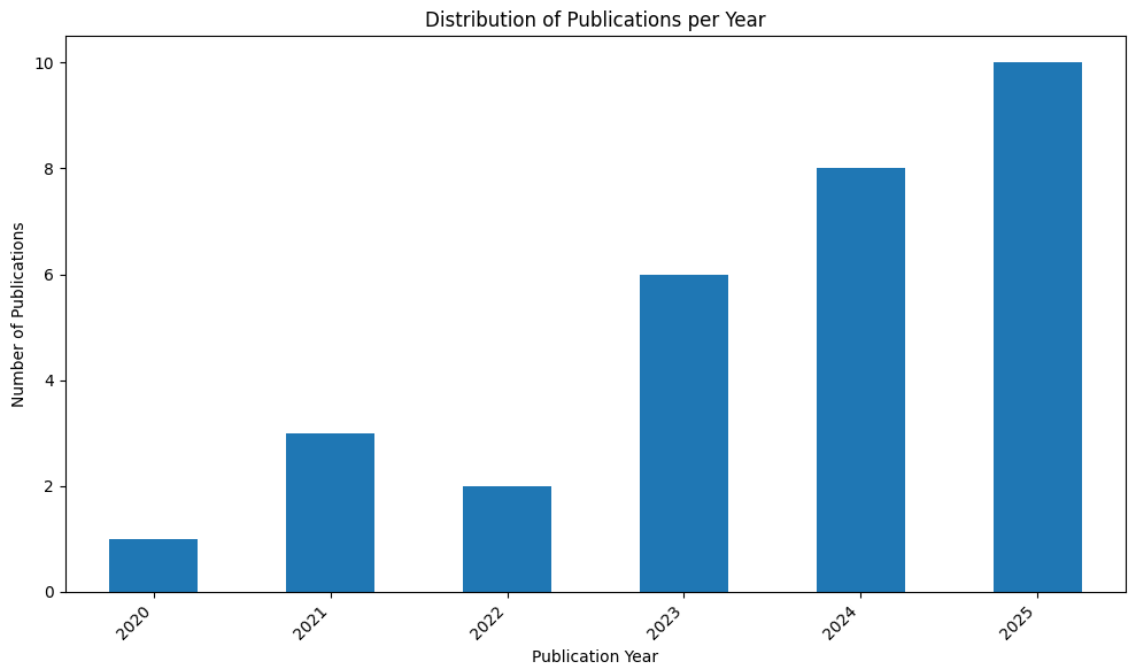
Setelah semua data dianalisis, langkah berikutnya adalah menginterpretasikan hasilnya untuk membuat cerita yang lengkap tentang kemajuan penelitian xG. Fokus interpretasi adalah menjawab beberapa pertanyaan penelitian utama. Di antaranya adalah untuk melacak tren publikasi setiap tahun dan menemukan bidang mana xG yang paling sering digunakan. Selain itu, dibahas metode pemodelan yang paling umum dan evolusinya. Selain itu, dibahas bagaimana peta konsep penelitian xG dibuat dengan menggunakan analisis jaringan kata kunci.

3. HASIL DAN PEMBAHASAN

Hasil analisis bibliometrik terhadap 30 (tiga puluh) artikel terpilih yang membahas pemodelan xG disajikan dan dibahas di bagian ini.

3.1. Tren Publikasi Tahunan

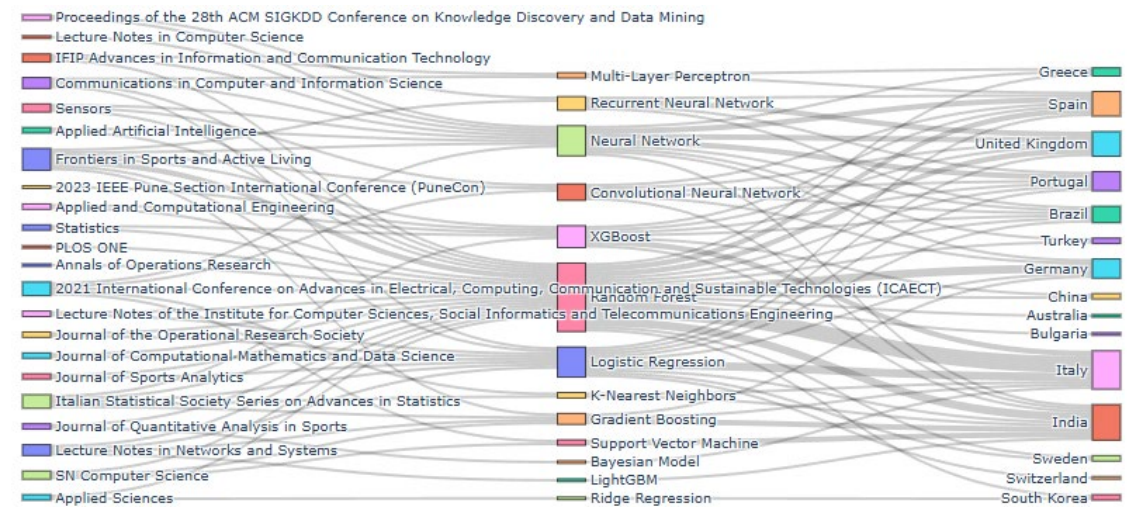
Analisis distribusi publikasi menunjukkan minat penelitian yang terus meningkat terhadap topik xG. Gambar 2 menunjukkan tren kenaikan yang signifikan dari tahun 2020 hingga 2025, meskipun ada sedikit penurunan pada tahun 2022, tren secara keseluruhan positif, dengan puncak tertinggi pada tahun 2025. Hal ini mengindikasikan bahwa xG bukan hanya sekadar metrik populer di kalangan media dan penggemar, tetapi juga telah menjadi subjek penelitian akademik yang matang dan terus berkembang.



Gambar 2 Distribusi Publikasi per Tahun 2020-2025

3.2. Analisis Demografi dan Kolaborasi Global

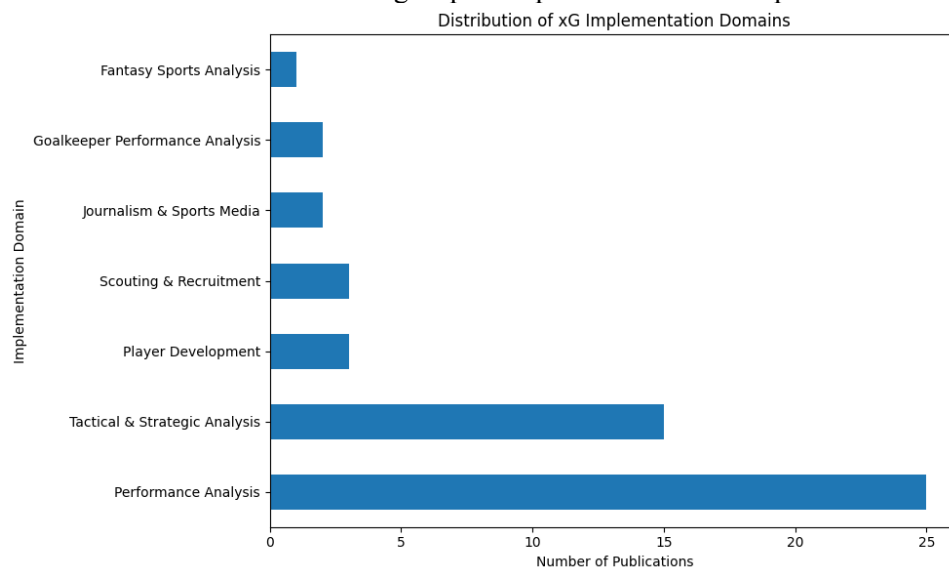
Untuk mendapatkan gambaran umum mengenai lanskap penelitian, diagram *Sankey* digunakan untuk memvisualisasikan hubungan tiga-arah antara sumber publikasi (jurnal/konferensi), metode/kata kunci yang digunakan, dan negara asal peneliti. Seperti terlihat pada Gambar 3, diagram ini menyoroti beberapa poin penting. Penelitian xG dipublikasikan di berbagai jenis sumber, mulai dari yang sangat teknis seperti *Lecture Notes in Computer Science* hingga jurnal spesialis olahraga seperti *Journal of Sports Analytics*, yang menandakan sifat interdisipliner dari topik ini. Dari sisi metode mulai dari "*Logistic Regression*" hingga "*XGBoost*" dan "*Neural Network*". Secara geografis, penelitian ini didominasi oleh peneliti dari negara-negara Eropa dengan tradisi sepak bola yang kuat seperti Spanyol, Portugal, Inggris, dan Italia, yang menunjukkan adanya ekosistem analisis olahraga yang matang di wilayah tersebut.



Gambar 3 Diagram *Sankey* Hubungan Sumber, Metode, dan Negara Peneliti

3.4. Domain Implementasi xG

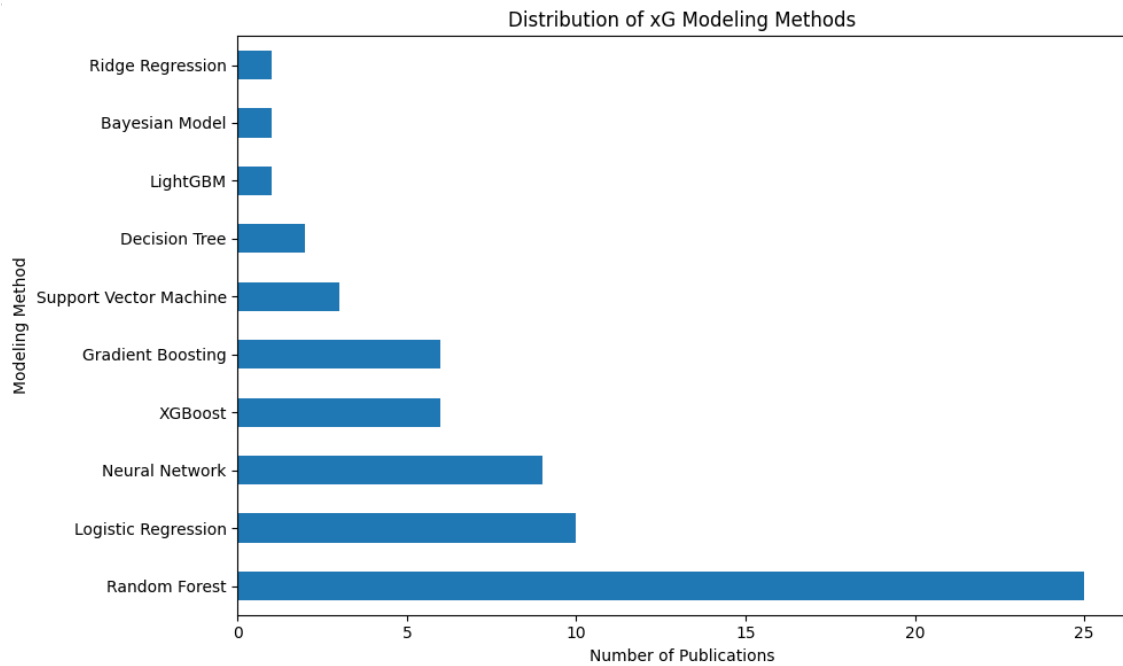
Analisis terhadap domain implementasi menunjukkan bahwa xG adalah metrik yang sangat fleksibel dan dapat diterapkan dalam berbagai konteks analisis sepak bola. Gambar 4 menunjukkan bahwa *Performance Analysis* adalah domain yang paling dominan, dengan 25 publikasi berfokus pada evaluasi pemain dan tim. Ini diikuti oleh *Tactical & Strategic Analysis* (15 publikasi), yang menggunakan xG untuk mengevaluasi efektivitas gaya bermain atau formasi. Domain lain yang juga signifikan adalah *Player Development* (3 publikasi), *Scouting & Recruitment* (3 publikasi), *Journalism & Sports Media* (2 publikasi), *Goalkeeper Performance Analysis* (2 publikasi), dan *Fantasy Sports Analysis* (1 publikasi). Sebaran ini menunjukkan difusi metrik xG dari ranah akademik ke berbagai aplikasi praktis di industri sepak bola.



Gambar 4 Distribusi Domain Implementasi xG

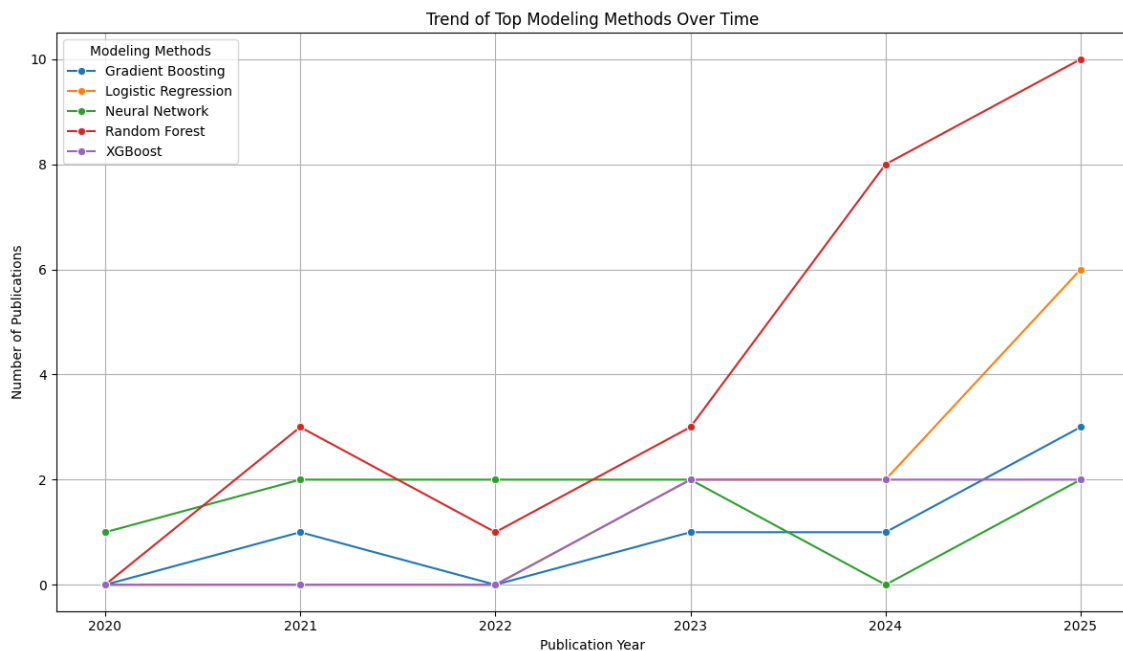
3.5. Analisis Metode Pemodelan

Salah satu fokus utama penelitian ini adalah mengidentifikasi metode pemodelan yang paling populer. Gambar 5 menunjukkan distribusi frekuensi dari berbagai metode yang digunakan. *Random Forest* muncul sebagai metode yang paling banyak digunakan (25 publikasi), mengungguli metode lain. Ini kemungkinan disebabkan oleh kemampuannya menangani interaksi kompleks antar variabel dan ketahanannya terhadap *overfitting*. Regresi Logistik dan *Neural Network* juga merupakan pilihan populer, masing-masing dengan 10 publikasi. Metode berbasis *boosting* seperti XGBoost dan *Gradient Boosting* juga menunjukkan kehadiran yang kuat, menandakan pergeseran ke arah model ensemble yang lebih canggih.



Gambar 5 Distribusi Metode Pemodelan xG

Untuk melihat evolusi tren ini dari waktu ke waktu, Gambar 6 memvisualisasikan penggunaan lima metode teratas setiap tahunnya. Grafik ini mengungkap tren yang menarik: sementara *Random Forest* menunjukkan peningkatan penggunaan yang stabil dan mencapai puncaknya pada tahun 2025, Regresi Logistik mengalami kebangkitan kembali popularitasnya pada tahun 2024 dan 2025. Ini mungkin menunjukkan bahwa meskipun model yang kompleks semakin populer, kesederhanaan dan interpretabilitas Regresi Logistik masih membuatnya relevan, terutama sebagai model dasar (*baseline*). *Neural Network* menunjukkan penggunaan yang fluktuatif, sementara *Gradient Boosting* dan *XGBoost* mulai mendapatkan daya tarik di tahun-tahun terakhir.



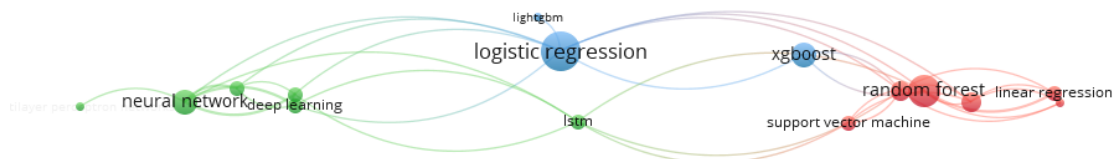
Gambar 6 Tren Metode Pemodelan Teratas dari 2020-2025

3.6. Analisis Jaringan dan Tren Kata Kunci

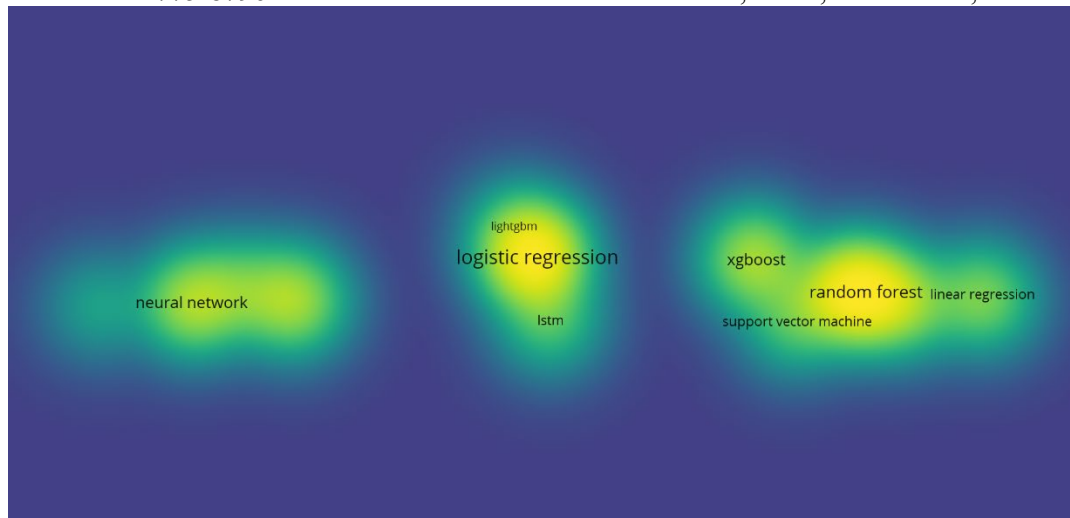
Analisis *co-occurrence* kata kunci menggunakan VOSviewer menghasilkan peta jaringan yang mengidentifikasi tiga kluster tematik utama dalam penelitian xG, seperti yang terlihat pada Gambar 7 dan 8. Penjelasan detail mengenai setiap kluster disajikan pada Tabel 2.

Tabel 2 Kluster Analisis *Co-Occurrence*

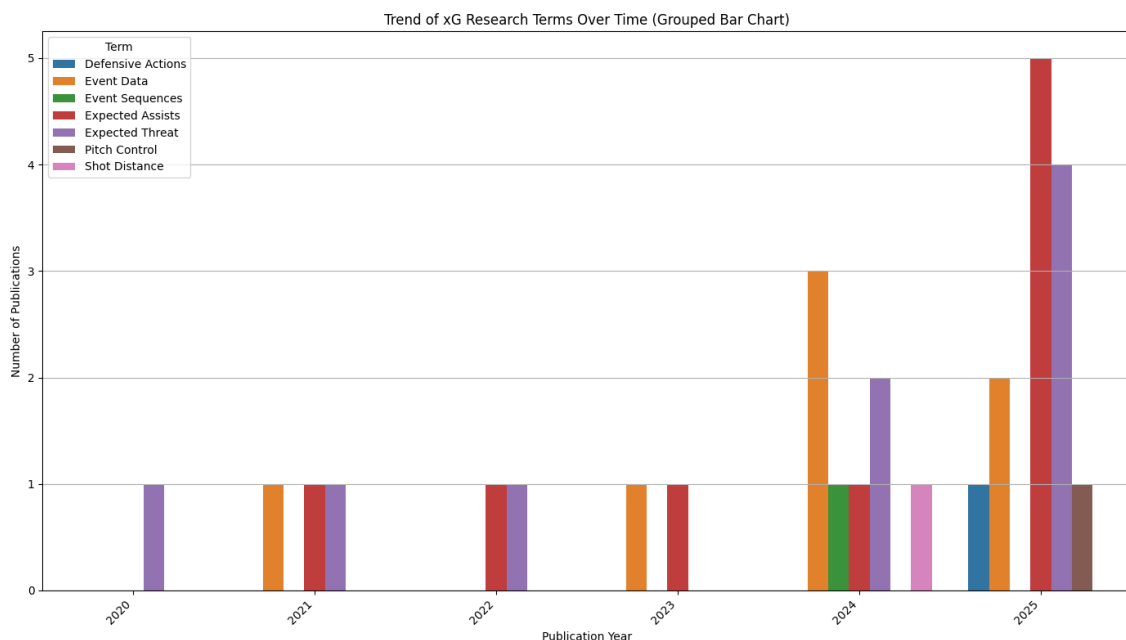
Kluster	Kata Kunci Utama	Deskripsi
Merah (<i>Machine Learning Klasik</i>)	<i>random forest, support vector machine, linear regression</i>	Mewakili fondasi metode machine learning yang sering digunakan untuk pemodelan xG.
Biru (<i>Model Regresi & Boosting</i>)	<i>logistic regression, xgboost</i>	Menghubungkan model statistik tradisional dengan teknik boosting modern.
Hijau (<i>Deep Learning</i>)	<i>neural network, deep learning</i>	Mewakili garda terdepan penelitian xG yang menggunakan arsitektur yang lebih kompleks.



Gambar 7 Analisis Jaringan *Co-occurrence Metode Pemodelan*

Gambar 8 Analisis *Density* Metode Pemodelan

Lebih lanjut, analisis tren istilah penelitian (Gambar 9) menunjukkan evolusi fokus dalam komunitas riset. Istilah-istilah dasar seperti *Event Data* dan *Shot Distance* hadir secara konsisten. Namun, pada tahun 2024 dan 2025, terjadi lonjakan signifikan pada istilah yang lebih canggih seperti *Expected Threat* (xT) [9] dan *Pitch Control* [10], yang merupakan pengembangan dari konsep xG untuk menilai aksi non-tembakan. Ini menandakan bahwa penelitian bergerak melampaui evaluasi tembakan semata menuju penilaian holistik terhadap semua aksi di lapangan yang menciptakan nilai.



Gambar 9 Tren Istilah Penelitian Terkait xG

4. KESIMPULAN

Analisis bibliometrik ini secara komprehensif memetakan lanskap penelitian pemodelan *Expected Goals* (xG) dari tahun 2020 hingga 2025. Kesimpulan utama dapat ditarik sebagai berikut:

Xxxx (xxxxxxxxxx)

- Minat akademik terhadap xG terus meningkat secara signifikan, menunjukkan relevansi dan pentingnya topik ini dalam analisis olahraga.
- *Random Forest* adalah metode pemodelan yang paling dominan, namun Regresi Logistik tetap relevan dan mengalami kebangkitan. Tren ini mengindikasikan adanya pergeseran ke arah model *machine learning* yang lebih kompleks, namun tanpa meninggalkan model dasar yang dapat diinterpretasikan.
- Aplikasi utama xG adalah untuk Analisis Performa, yang menegaskan perannya sebagai alat evaluasi fundamental bagi pemain dan tim.
- Penelitian mulai bergerak melampaui xG itu sendiri, dengan meningkatnya minat pada metrik turunan yang lebih canggih seperti *Expected Threat* dan *Pitch Control*.

Kelebihan studi ini adalah penyajian visual yang jelas mengenai tren dan hubungan dalam literatur xG. Namun, kekurangannya adalah cakupan data yang terbatas pada basis data Scopus dan fokus pada analisis kuantitatif tanpa mengevaluasi kualitas atau akurasi dari model-model yang dibahas.

5. SARAN

Berdasarkan temuan dan keterbatasan penelitian, beberapa saran untuk penelitian di masa depan dapat diajukan:

- Penelitian selanjutnya dapat melakukan meta-analisis untuk membandingkan secara kuantitatif performa dari berbagai metode pemodelan xG yang telah diidentifikasi populer dalam studi ini.
- Disarankan untuk memperluas analisis dengan memasukkan data dari sumber lain seperti Dimensions.ai, *Web of Science*, atau *Google Scholar* untuk mendapatkan gambaran yang lebih lengkap.
- Fokus penelitian masa depan dapat diarahkan pada tren penggunaan fitur (*feature*) dalam model xG, terutama yang berasal dari data pelacakan spasial-temporal (*tracking data*), seiring dengan meningkatnya ketersediaan data tersebut.

DAFTAR PUSTAKA

- [1] D. Sumpter, 2016, *Soccermatics: Mathematical Adventures in the Beautiful Game*, Bloomsbury Sigma, London.
- [2] W. Spearman, 2018, Beyond Expected Goals, di *12th Annual MIT Sloan Sports Analytics Conference*, Boston.
- [3] S. J. Brillinger, 2019, Assessing the Association Between the Expected Goals and the Actual Goals Scored in the English Premier League, *Journal of Sports Analytics*, vol. 5, no. 3, hlm 187-198.
- [4] S. Macdonald, 2012, An Introduction to Expected Goals in Football, *Journal of Quantitative Analysis in Sports*, vol. 8, no. 3.
- [5] P. Robberechts, J. Davis, dan J. Van Haaren, 2020, A Bayesian Approach to Expected Goals in Soccer, di *ECML PKDD 2019: Machine Learning and Knowledge Discovery in Databases*, hlm 250-265.
- [6] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, dan W. M. Lim, 2021, How to conduct a bibliometric analysis: An overview and guidelines, *Journal of Business Research*, vol. 133, hlm. 285-296.

- [7] G. Anzer, dan P. Bauer, 2021, A goal scoring probability model for shots in football, *Frontiers in Sports and Active Living*, vol 3, hlm 1-11.
- [8] S. Harish, A. A. Kevin, U. H. Vardhan, dan P. S. Femi, 2023, Expected Goals Prediction in Football using XGBoost, *ESP Journal of Engineering & Technology Advancements*, vol 3, no. 1, hlm 21-26.
- [9] K. Singh, 2019, *Introducing Expected Threat (xT)*, [Online]. Tersedia: <https://karun.in/blog/expected-threat.html>. [Diakses: 28-Agu-2025].
- [10] W. Spearman, 2018, A framework for tactical analysis and decision making in soccer using spatiotemporal data, *di 12th Annual MIT Sloan Sports Analytics Conference*, Boston.