Investigating UK Software Engineering Demand: Industry Distributions and Programming Language Trends

1. Introduction

The United Kingdom remains one of the world's leading destinations for software engineering employment, drawing a substantial number of developers and data analysts each year (Fearns et al., 2023). This prominence arises from the UK's vibrant and diverse technology ecosystem, where numerous industries—such as FinTech, software development, banking, retail, and telecommunications—compete to attract skilled software engineers. For aspiring developers entering the job market, understanding exactly which industries employ the largest number of software engineers, and identifying the most sought-after technical skills within those industries, becomes an essential step towards making informed career choices.

As an MSc student currently entering the software engineering job market, my motivation for this analysis is both personal and practical. Clearly identifying which industries employ the greatest number of software engineers, and which programming languages are most in demand within these industries, is critical to strategically positioning myself—and potentially other recent graduates—for successful employment outcomes.

Previous studies have indicated that traditional programming languages such as Java, Python, and C++ remain highly popular due to their wide application in enterprise solutions, web development, and data analytics (Patel and Tere, 2025). Meanwhile, emerging languages like TypeScript and Go are increasingly sought for their adaptability in modern technology infrastructures. Nonetheless, current empirical insights on the specific programming languages and skillsets demanded by each UK industry remain relatively limited. Given this context, this study is guided by the following research questions:

- Which industries in the UK currently have the highest employment of software engineers?
- Within these industries, which programming languages are in greatest demand?

To answer these questions, this study analyses data from the Stack Overflow Developer Survey 2023, specifically examining respondents who identified themselves as software developers based in the UK. By utilising both supervised (Random Forest Classification) and unsupervised (K-means clustering) machine learning techniques, this research provides empirical insights into industry distributions and programming language preferences. Ultimately, the findings aim to equip prospective software engineers with strategic information that could enhance their employability and guide their ongoing professional development.

2. Literature Review

Software engineering skills are crucial to the UK's technological innovation and economic growth. Harrison (2012) highlights the importance of aligning educational curricula with market demands, arguing this alignment directly enhances national competitiveness. Gurcan and Köse (2017) further stress that software engineers must not only master current technologies but also adapt rapidly to emerging tools and industry changes.

Programming language trends have shifted significantly in recent years. Traditional languages like Java, C++, and Python remain critical due to their extensive infrastructures and developer communities (Chen et al., 2005). However, languages such as Python and JavaScript are increasingly popular, driven by their adaptability to cloud computing, data science, and machine learning contexts (Patel and Tere, 2025). Additionally, emerging languages like TypeScript and Go are gaining industry traction due to performance and scalability advantages (Patel and Tere, 2025).

Patel and Tere (2025) also identify clear differences in language adoption across industries: the financial and Fintech sectors heavily utilise Java and SQL for robust backend systems, whereas retail and consumer-focused industries prefer flexible, web-oriented languages like JavaScript. These findings underline the necessity for developers to strategically select skills aligning with targeted career sectors.

Educational institutions must continually adapt curricula to reflect these evolving industry needs. Persistent mismatches between academic training and industry requirements exacerbate skill gaps, limiting graduates' immediate employability (Harrison, 2012; Gurcan and Köse, 2017).

Therefore, clear empirical analysis of programming language demand across UK industries provides strategic guidance for graduates aiming to enter the software development market, directly informing this study's aims and analytical focus (Fearns et al., 2023).

3. Data processing

3.1 Data Source and Description

The primary dataset employed in this study is drawn from the Stack Overflow Developer Survey 2023, a recognised annual survey capturing global software development trends. To focus on the UK context, data were filtered to include only respondents who indicated their location as "United Kingdom of Great Britain and Northern Ireland," yielding 3,224 valid entries. While this survey may not represent every software professional in the UK, it provides a valuable snapshot of key employment sectors and programming language usage (Fearns et al., 2023).

3.2 Missing Data Handling

The initial exploratory data analysis identified significant missingness (50.81%) in the "Industry" field. Completely removing all respondents with missing industry information would drastically reduce the dataset size and potentially bias the analysis. To mitigate this, missing industry data was imputed by assigning the category "Unknown". This step ensured retention of the maximum number of respondents. However, when conducting detailed industry-specific analyses, responses labelled as "Unknown" were excluded to preserve analytical clarity and validity.

In contrast, the "LanguageHaveWorkedWith" field showed a much lower missingness rate (1.55%), allowing direct removal of these incomplete responses without substantially affecting the dataset's statistical integrity.

3.3 Data Cleaning and Transformation

To ensure consistency, leading and trailing whitespace was removed from the "Industry" column. Meanwhile, the "LanguageHaveWorkedWith" column—originally storing multiple languages in a semicolon-separated format—was split into discrete strings to accurately count the usage frequency of each language. Once extracted, this language data was converted into binary indicators (one-hot encoding). Each respondent thus had a vector of language features, where a value of 1 indicated usage of that language, and 0 indicated non-usage. This approach standardised the data for subsequent statistical and machine learning tasks.

3.4 Feature Selection for Modelling

For classification tasks, binary-encoding of the top programming languages was necessary. Rows were converted into a one-hot encoded Data Frame, where each column indicated whether a respondent worked with a particular programming language. Maintaining the top 10 languages helped reduce sparsity while retaining the languages most widely reported.

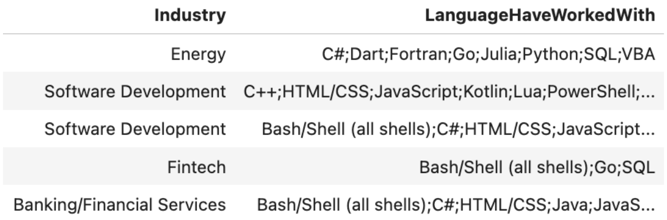
By combining frequency counts, sector filtering, and standard text-cleaning approaches, the dataset provided a clearer foundation for subsequent analytics on industry distribution and language usage among UK-based developers.

3.5 Analytic Methods

Following the above data processing steps, two analytic stages were undertaken:

1. Exploratory Data Analysis (EDA):
   To identify industries employing the largest number of software engineers, simple frequency counts were calculated and visualised using horizontal bar charts to illustrate clearly the ranking of industries based on respondent frequency.
2. Frequency and Visualisation Analysis:
   Programming language popularity was assessed by counting occurrences within the pre-processed dataset, resulting in clear visual representation through bar charts. These charts highlighted the most frequently used languages within the top three identified industries.

Further analysis, including clustering or regression modelling, could be applied in future studies but is outside the scope of this initial investigation. The chosen methodologies provide clear and accessible insights suitable for informing prospective job seekers and aligning educational focus with market needs.



Industry	LanguageHaveWorkedWith
Energy	C#;Dart;Fortran;Go;Julia;Python;SQL;VBA
Software Development	C++;HTML/CSS;JavaScript;Kotlin;Lua;PowerShell;...
Software Development	Bash/Shell (all shells);C#;HTML/CSS;JavaScript...
Fintech	Bash/Shell (all shells);Go;SQL
Banking/Financial Services	Bash/Shell (all shells);C#;HTML/CSS;Java;JavaS...

Figure 1: Sample of the UK Dataset (Stack Overflow Developer Survey 2023)

4. Learning Methods

To address the research objectives effectively, two distinct data analytic methods were selected and applied: Random Forest Classification (a supervised learning approach) and K-means Clustering (an unsupervised method).

4.1Random Forest Classification

Random Forest is a supervised machine learning algorithm that builds multiple decision trees by repeatedly sampling subsets of the data and aggregates their predictions to deliver stable and reliable classification results.

Justification for selecting Random Forest:

- Robustness to complex, non-linear relationships:
  Given that developers typically use multiple languages simultaneously, Random Forest effectively captures potential complex interactions among language features.
- Effective handling of imbalanced and heterogeneous data:
  With imbalanced industry distributions (e.g., Software Development being the dominant category), Random Forest provides mechanisms like class weighting (class_weight='balanced') to reduce bias towards majority classes.
- Clear Feature Importance Ranking:
  Random Forest clearly ranks features by their predictive strength, providing explicit insights into which programming languages best predict industry affiliation—crucial information for job seekers.

Application and Validation of Random Forest

The prepared dataset, consisting of binary-encoded top 10 programming languages as features (X) and industry classifications as the target (y), was split into a training set (80%) and a test set (20%). To ensure robustness, 5-fold cross-validation was utilised to tune hyperparameters (e.g., the number of estimators and maximum depth). The optimal model was then tested on the unseen hold-out set, with accuracy, precision, recall, and a confusion matrix reported to clearly document model strengths and limitations.

4.2 K-means Clustering

Complementing the supervised analysis, K-means clustering was employed as an unsupervised learning method to reveal intrinsic patterns within programming language usage across industries.

Justification for selecting K-means clustering:

- Discovery of natural skill-based groupings:
  K-means clustering effectively identifies clusters based solely on language skills, revealing sectors or roles sharing similar skill demands, thus helping job seekers understand transferable career opportunities.
- Interpretability and simplicity:
  The simplicity of K-means clustering facilitates intuitive interpretation of resulting groups, making it accessible and meaningful for students and industry stakeholders seeking practical insights.
- Scalability and computational efficiency:
  K-means efficiently handles large datasets, allowing for quick iteration and experimentation to find meaningful clusters.

Application of K-means clustering

The same binary-encoded language matrix was utilised for clustering. The optimal number of clusters (k) was determined by applying the Elbow Method, testing k values between 3 and 6. k=5 was ultimately selected, as it provided interpretable clusters that highlighted meaningful distinctions in language usage, balancing interpretability and granularity.

By combining these supervised and unsupervised approaches, this study offers comprehensive insights: the Random Forest classification helps quantify specific programming languages' predictive importance for industry roles, while K-means clustering uncovers broader skillset groupings relevant to multiple industries.

5. Analysis & Results

5.1 Exploratory Data Analysis: Industry Distribution

The initial exploratory data analysis (EDA) involved identifying which UK industries currently employ the largest number of software developers. After excluding respondents labelled as ‘Unknown’ (due to missing industry data), the analysis covered 2,372 valid respondents. Figure 2 presents the ten most frequent industries identified:

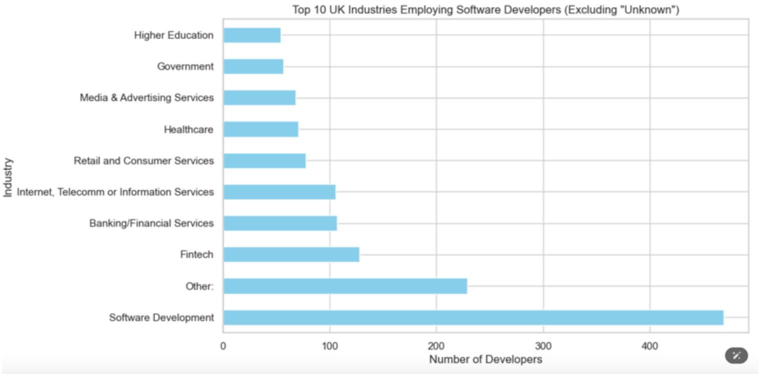


Figure 2: Top 10 UK Industries Employing Software Developers (Stack Overflow 2023)

As Figure 2 demonstrates, Software Development clearly dominates the dataset, followed by sectors like Fintech, Banking/Financial Services, and Internet and Telecommunications. These findings align with Patel and Tere (2025), who similarly identified strong software engineering demand in finance-related and software-intensive sectors within the UK.

5.1.1 Analysis of Programming Language Usage

Programming languages were further analysed to identify those most frequently demanded across the top UK industries. Figure 3 illustrates the top 10 programming languages reported by UK-based software developers.

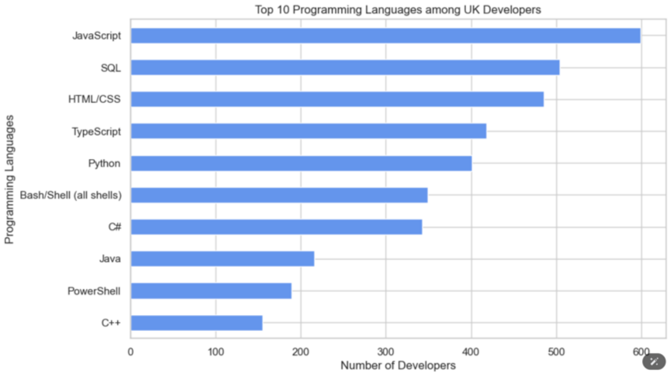


Figure 3: Top 10 Programming Languages among UK Developers (Stack Overflow 2023)

From Figure 3, JavaScript, SQL, HTML/CSS, TypeScript, and Python are prominently used among developers across industries. These languages consistently rank highly, confirming their broad relevance to software engineering roles in the UK. The prevalence of JavaScript and HTML/CSS reflects the continued importance of web development skills, while SQL underscores the demand for data management and backend system expertise.

To provide deeper insights into language preferences within individual industry sectors, Figure 4 compares the top five programming languages within each of the five major UK industries identified previously.

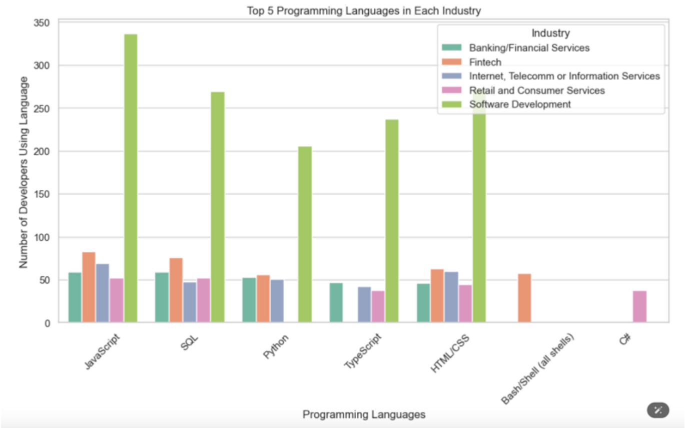


Figure 4: Top 5 Programming Languages in Each Major UK Industry (Stack Overflow 2023)

Figure 4 provides a comparative overview of the five most frequently reported programming languages within each of the top five UK industries. Notably, clear patterns emerge, highlighting both industry-specific trends and overlapping skill requirements.

- Software Development:
  Shows high usage across all major languages (JavaScript, HTML/CSS, Python, SQL, TypeScript), indicating roles typically require versatile skills, spanning both frontend and backend development.
- Banking and Financial Services:
  Strong focus on SQL, highlighting backend database management needs. Moderate use of JavaScript and Python suggests roles often combine secure data operations with customer-facing applications.
- Fintech:
  Heavily relies on SQL for data-intensive tasks like financial analytics. Python, JavaScript, and notable Bash/Shell usage indicate additional requirements for analytical tools, customer interfaces, and automation.
- Internet, Telecommunications, and Information Services:
  Primarily uses JavaScript and HTML/CSS, reflecting web-centric services. Python is also prominent for backend and analytics tasks, while TypeScript is less common.
- Retail and Consumer Services:
  Balanced usage with emphasis on JavaScript, SQL, and notably C#, suggesting preference for robust enterprise-level backend systems managing customer data and inventory.

Overall, the analysis demonstrates distinct industry-specific preferences alongside considerable skill overlap, particularly in web (JavaScript, HTML/CSS) and data-oriented languages (SQL, Python). For job seekers and new graduates, proficiency in these broadly applicable languages can significantly enhance cross-industry employability, whereas familiarity with specialised tools such as Bash/Shell and C# can provide additional advantages in specific sectors.

5.2 Random Forest Classification Analysis and Results

The final Random Forest model achieved an overall classification accuracy of approximately 41% on the held-out test set. Table 1 summarises detailed classification metrics, including precision, recall, and F1-scores for each industry category:

Industry	Precision	Recall	F1-score	Support
Software Development	0.58	0.62	0.60	97
Banking/Financial Services	0.19	0.22	0.21	18
Fintech	0.11	0.07	0.09	29
Internet, Telecom or Information Services	0.33	0.22	0.26	23
Retail and Consumer Services	0.10	0.18	0.12	11
Accuracy	41%	-	-	178

Table 1: Random Forest Classification Report

The confusion matrix showed that the model performed best in identifying "Software Development," due to the larger volume of training examples. However, performance was weaker in smaller categories like FinTech, highlighting the difficulty of accurately classifying less-represented industries.

Feature Importance

An important advantage of Random Forest is featuring importance rankings. The results indicated that TypeScript, Python, and JavaScript were the most influential predictors. The high importance of these languages suggests their strong correlation with specific industry sectors (e.g., TypeScript and JavaScript predominantly for web-focused roles, Java and SQL for backend financial services).

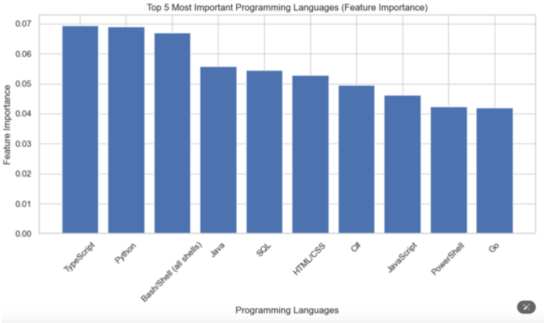


Figure 5: Top 5 Programming Languages by Feature Importance (Random Forest)

Strengths and Weaknesses of Random Forest

Strengths:

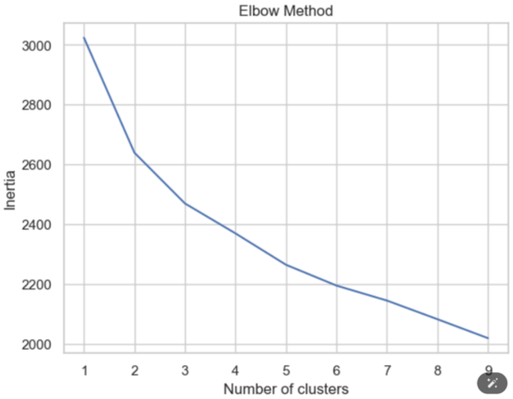
- Effectively captures complex, non-linear relationships between programming languages and industry classifications, accurately reflecting real-world data complexities.
- Clearly identifies feature importance, offering actionable insights into which programming languages are most predictive of employment sectors, thus benefiting career-focused decision-making.

Weaknesses:

- Moderate overall classification accuracy, primarily due to substantial class imbalance and overlapping language usage patterns across industries.
- Limited predictive accuracy in smaller, less represented industry sectors.

5.3 K-means Clustering Results

Complementing supervised results, K-means clustering (k=5) (figure 6) was applied to reveal natural skill groupings. Figure 7 illustrates the distribution of industries across these five clusters:



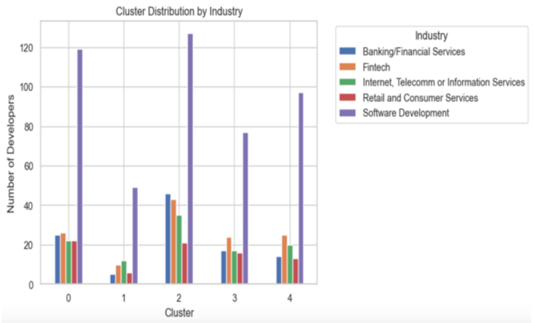


Figure 6: Determining Optimal Number of Clusters Using the Elbow Method

Figure 7: Cluster Distribution by Industry

Distinct clusters emerged:

- Cluster 0: Dominated by Software Development, reflecting broad generalist language use.
- Cluster 1: Smallest cluster; no clear dominant industry—suggesting niche or less clearly-defined roles.
- Cluster 2: Large, diverse cluster; frequent in Banking and Fintech, suggesting roles heavily focused on SQL and Python.
- Clusters 3 & 4: Moderate-sized clusters primarily representing Software Development but also balanced across other sectors, indicating a diverse and transferable skillset among developers.

Strengths and Weaknesses of K-means Clustering

Strengths:

- Clearly identifies cross-industry skill clusters, providing practical career insights.
- Simple and interpretable cluster structures aiding career strategy formulation.

Weaknesses:

- Clusters lack nuanced differentiation for niche roles or specific technologies.
- Choice of k=5, while balanced, may obscure finer distinctions within large clusters.

6. Concluding Remarks

6.1 Summary and Achievements

This study systematically analysed UK industry employment trends and programming language demands, leveraging data from the Stack Overflow Developer Survey 2023. Through Exploratory Data Analysis (EDA), Random Forest Classification, and K-means clustering, several significant insights were achieved:

- Clear identification of the leading UK industries for software developers, with Software Development, Banking and Financial Services, and Fintech emerging as the most significant sectors.
- Programming language analysis demonstrated that JavaScript, SQL, HTML/CSS, TypeScript, and Python consistently ranked highly, suggesting broad cross-industry applicability of these languages.
- Supervised classification using Random Forest successfully identified programming languages predictive of industry classification, particularly within "Software Development"; notably, JavaScript, despite widespread use, lacked predictive importance due to its ubiquity across sectors.
- Unsupervised K-means clustering revealed natural groupings of developer skills, indicating clear overlaps in programming language usage across industries. This finding provided valuable insights into potential skill transferability between industries.

These insights provide valuable strategic direction for prospective software developers regarding essential languages to prioritise during training and career planning.

6.2 Limitations

Despite clear insights, several limitations emerged during analysis:

- Class Imbalance: The Random Forest model encountered performance degradation due to imbalanced class distributions, where some industries (e.g., Software Development) greatly outnumbered others (e.g., Fintech, Retail). This imbalance negatively impacted the predictive accuracy for smaller categories.
- Limited granularity of clustering: The unsupervised K-means clustering approach effectively identified broad skill clusters but could not precisely differentiate finer specialisations or niche technical roles within each industry cluster.
- Dataset Representativeness: Reliance on Stack Overflow survey data might introduce respondent bias, as not all UK software engineers necessarily engage with or respond to such surveys, potentially affecting representativeness.

6.3 Future Improvements and Research Directions

To address these limitations, future research could:

- Utilise oversampling methods (e.g., SMOTE) to enhance Random Forest accuracy for minority industry classifications, balancing class distributions more effectively.
- Experiment with alternative clustering methods (e.g., hierarchical clustering, DBSCAN) to provide finer distinctions within clusters, enhancing the interpretability of skill-based industry grouping.
- Expand the dataset by incorporating other complementary data sources, such as industry-specific surveys or employment databases, to strengthen representativeness and mitigate selection bias.

6.4 Conclusion

This project successfully identified the major UK industry sectors employing software engineers and highlighted programming languages with the greatest demand across these industries. It illustrated that broad, versatile skills in languages such as JavaScript, Python, and SQL significantly enhance employability due to their cross-industry relevance, while also pinpointing language combinations beneficial for more specialised roles in sectors like Banking or Retail.

Despite methodological limitations and moderate classification performance, the combination of Random Forest and K-means clustering delivered practical insights valuable to students, educators, and industry stakeholders. Ultimately, the results underline the necessity of educational institutions to continually adapt curricula to align closely with dynamic, industry-specific skill requirements, thereby improving graduates' employability and career outcomes.

7. References

Stack Overflow Developer Survey 2023 (2024) Stack Overflow Annual Developer Survey 2023. Available from: https://insights.stackoverflow.com/survey [Accessed 1 March 2025].

Fearns, J., Harriss, L. and Lally, C. (2023) Data science skills in the UK workforce. UK Parliament POSTnote No. 697. Available at: https://researchbriefings.files.parliament.uk/documents/POST-PN-0697/POST-PN-0697.pdf [Accessed: 3 March 2025].

Patel, S. and Tere, G. (2025) Analyzing Programming Language Trends Across Industries: Adoption Patterns and Future Directions. International journal of emerging science and engineering. [Online] 13 (2), 19–26. [Accessed: 7 March 2025].

Gurcan, F. and Köse, C. (2017) Analysis of software engineering industry needs and trends: implications for education. International Journal of Engineering Education. 33 (4), 1361–1368. [online]. Available from: https://dialnet.unirioja.es/servlet/articulo?codigo=6897054. [Accessed: 6 March 2025].

Chen, Y., Dios, R., Mili, A., Wu, L. and Wang, K. (2005) An empirical study of programming language trends. IEEE Software. [Online] 22 (3), 72–78. [online]. Available from: https://www.computer.org/csdl/magazine/so/2005/03/s3072/13rRUwgyOep. [Accessed: 6 March 2025].

Thompson, J. B. and Stobart, S. C. (1993) Software engineering in the commercial sector present and future: A United Kingdom perspective. Computer Software and Applications Conference. [Online] 76–82. [Accessed: 8 March 2025].

Harrison, M. (2012) Jobs and growth: The importance of engineering skills to the UK economy, Royal Academy of Engineering Report, pp. 1–35. [online]. Available from: https://www.voced.edu.au/content/ngv%3A53287. [Accessed: 8 March 2025].

Edwards, H. (1999). Software engineering education from a UK academic’s perspective. Computer Software and Applications Conference. https://doi.org/10.1109/CMPSAC.1999.812709 [Accessed: 9 March 2025].