Article

Integrated Survival Model for Predicting Patent Litigation Hazard

Youngho Kim ¹ , Sangsung Park ² , Junseok Lee ³, Dongsik Jang ¹ and Jiho Kang ^{4,*}¹ Department of Industrial Management Engineering, Korea University, Seoul 02841, Korea; youngho0928@korea.ac.kr (Y.K.); jang@korea.ac.kr (D.J.)² Department of Big Data and Statistics, Cheongju University, Chungbuk 28503, Korea; hanyul@cju.ac.kr³ MICUBE Solution, Seoul 06719, Korea; js.lee@micube.co.kr⁴ Machine Learning Big Data Institute, Korea University, Seoul 02841, Korea

* Correspondence: kangmae@korea.ac.kr

Abstract: Patent litigation occurs when a company's product or service violates the scope of another company's patent rights. When they occur, companies suffer a disruption to the sales of their products and services, thus hindering the sustainability of their business activities. For this reason, companies have established and analyzed wide-ranging strategies to prevent patent litigation. Of those, statistical and machine learning-based quantitative methods using patent big data have several advantages, such as a reduced cost and objective results. Existing quantitative methods analyze patent information and litigation based on the time of data collection. However, the values of patents and their litigation hazards change over time. In addition, the existing methods do not take into account censored data; that is, patents that may result in litigation after the data is collected. In this paper, to solve this problem we propose an integrated survival model that considers censored data and predicts patent litigation hazards over time. The proposed model is a non-parametric survival analysis method based on a random survival forest. It uses pre-trained word2vec and clustering to effectively reflect the technology fields as well as the quantitative information of the patent. The word2vec is a technique for natural language processing and enables the use of patent text information. In order to examine the practicality of the integrated survival model, an experiment is conducted with patent big data related to sensor semiconductors based on AI technology applicable to robotics. In the experiment, it was found that the litigation hazard occurred 150 months after the patent application and increase rapidly from 200 months. Furthermore, the proposed model showed better predictive performance than other survival analysis models. The proposed model could be used by potential defendants to protect their patents.

Citation: Kim, Y.; Park, S.; Lee, J.; Jang, D.; Kang, J. Integrated Survival Model for Predicting Patent Litigation Hazard. *Sustainability* **2021**, *13*, 1763. <https://doi.org/10.3390/su13041763>

Academic Editor: Alan Randall

Received: 14 December 2020

Accepted: 4 February 2021

Published: 6 February 2021

Keywords: patent litigation; survival analysis; patent big data; text mining; random survival forest

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Patents contain varied and detailed information about the developed technology and indicate exclusive rights [1,2]. Companies can realize profit-making through patents, and the patents legally protect the technology the companies have developed [3]. For this reason, the importance of patents has been increasingly drawing attention, and the number of global patent applications is rising [4]. Additionally, as patents have become big data in recent years, the term patent big data is often used. Patent litigation occurs when a company's patents, products or other operations infringe on the scope of other companies' patent rights. When patent litigation occurs, companies' sales of products and services are disrupted. In addition, much time and resources have to be invested into resolving disputes, and this interferes with the sustainability of management activities and causes a decline in competitiveness [5,6]. Therefore, many companies perform licensing, portfolio strategies, and competitor analysis to prevent patent litigation [7–11]. Among them, quantitative methods that use statistics or machine learning based on a vast volume

of patent big data can reduce costs and ensure the objectivity of the derived results. Thus, it is possible to construct a sustainable patent litigation prediction model using quantitative methods based on patent big data.

Existing quantitative methods for patent litigation analysis reflect patent information and litigation status in model construction based on the time of data collection [12–21]. Therefore, patent information and litigation hazards that change over time cannot be considered. In this paper, we propose a sustainable model that can predict patent litigation hazards over time based on the random survival forest (RSF). The proposed model applies pre-trained word2vec to patent text information (TI) and performs K-means clustering so that not only quantitative information (QI) but also technology fields can be effectively reflected in the model. The proposed model differs from the existing RSF model because it combines natural language processing techniques and clustering to use TI. This paper is organized as follows. Section 2 explains the theories and related studies necessary to build an integrated survival model. Section 3 describes in detail the construction process of the proposed model. Section 4 presents an experiment conducted to verify the practical applicability of the integrated survival analysis and its results. Section 5 discusses the proposed model and its limitations, and Section 6 provides conclusions and suggestions for future research directions.

2. Related Work

2.1. Quantitative Methods for Patent Litigation Analysis

Previous studies extracted QI and TI from collected patent data to search for litigation occurrence factors or construct prediction model [12–21]. Lai and Che (2009) proposed a quantitative method to predict the legal value of a patent [12]. They extracted QI from litigation patents and used the Kaiser-Meyer-Olkin (KMO) test to identify important factors. The derived factors were used for neural network training to construct a dispute amount prediction model. The study of Lai and Che (2009) is applicable to patents for which litigation has occurred, but has limitations for predicting litigation hazards in advance. Chien (2011) conducted a study on the construction of a patent litigation prediction model using logistic regression [13]. Chien (2011) collected litigation patents and non-litigation patents, respectively, extracted QI, and trained the model. Alex et al. (2002) extracted QI from collected patent data and identified important variables through stepwise selection [14]. Additionally, the logistic regression-based patent litigation prediction model was proposed using only the identified important variables. Chien (2011) and Alex et al. (2002) extracted QI for not only the patents over which litigation had occurred, but also those for which it had not, and used it to construct a predictive model. However, it needs to be considered because the QI value is different depending on the technology fields of the patent and the frequency of litigation. Lanjouw (2001), Bessen and Meurer (2013) conducted a study to quantitatively analyze the major characteristics of patents over which litigation occurs [15,16]. They used various information such as patent QI, the technology fields involved, finance, and the type of applicant. Moreover, Lanjouw (2001), Bessen and Meurer (2013) discovered that the technology or industry fields to which the patent belongs affects litigation likelihoods. Lanjouw (2001) used the International Patent Classification (IPC) code granted to individual patents to identify technology fields. However, since the IPC code identifies technology fields according to eight previously classified sections, its application to convergence and new technologies is limited. Bessen and Meurer (2013) classified industry fields through companies' Standard Industry Classification (SIC) code assigned by Compustat, but SIC codes represent the fields of industry of the applicants themselves rather than those of the individual patent units, limiting accuracy. For this reason, some studies have proposed a method of effectively identifying a technology field using a patent TI that includes detailed information on the developed technology [17,18]. Kim (2014) used the ensemble model to effectively reflect the sparse characteristics of data when constructing a patent litigation prediction model [19]. As well as patent information, the model proposed by Kim (2014) additionally collects and considers information such as

technology transfer and multi-litigation. Cowart et al. (2014) compared logistic regression and classification trees to predict patent litigation [20]. Cowart et al. (2014) evaluated the quantitative model in various aspects such as the method of presenting the results and the type of input variables. Wongchaisuwat et al. (2016) proposed a method for predicting the litigation likelihood and litigation time [21]. They predicted litigation patents by combining the convex hull distance method and the ensemble classification model. They also proposed a model that divides the timing of patent litigation occurrences into three types and classifies them into hierarchical trees.

The above prior studies utilize various patent information and classification models to predict patent litigation. However, while patent information and litigation status are reflected in model construction based on the time of data collection, patents change in value and litigation hazard over time. For example, the citation and family indicators of a patent accumulate over time. Prior studies have built models based on the time of collection of patents, so they do not reflect patents that may result in litigation in the future, that is, censored data. For this reason, it is necessary to construct a model that reflects censored data and predicts patent litigation hazards over time. In this paper, we propose an integrated survival model that can predict patent litigation hazards over time based on the random survival forest.

2.2. Patent Survival Analysis

Survival analysis is a method of analyzing events and times of interest [22–25]. Survival analysis can be applied to data in which events occur during observation or collection periods. In most cases, not all events of interest occur during the observation of the data. Some data may also be incomplete due to other events or factors. As such, data for which the occurrence of an event cannot be known at the time the observation is completed is called censored [22]. General analysis methods treat censored data as missing values. However, in survival analysis, the censored data is considered important and included [26].

For patents, legal status changes such as applications and registrations are clear. Furthermore, censored data exists in patents for reasons such as application withdrawals and end-of-period expirations. For this reason, various existing studies have applied survival analysis to patent data [27–34]. Wagner and Cockburn (2010) analyzed the relationship between patent and company survival [27]. Wagner and Cockburn (2010) used the number of patents owned by a company as an input variable for the survival model. Zeebroeck (2007), Xie and Giles (2011), and Nikzad (2011) analyzed the time and influencing factors from patent application to approval using survival analysis [28–30]. They used QI as an input variable and applied it to parametric and non-parametric survival models. It was determined that the time required to move from application to approval varies according to country and technology field. Nakata and Zhang (2012) used a survival regression model to analyze the time of patent examination requests by Japanese electrical and electronic manufacturers [31]. Nakata and Zhang (2012) made measurable indicators of a company's patent application strategy and used it as an input variable along with QI. Han and Sohn (2015) used Weibull regression to evaluate patent value and applied text mining and singular value decomposition (SVD) to the patent TI and used it in the model [32]. Erzurumlu and Pachamanova (2020) used topic modeling and cox proportional hazard regression to predict the commercial viability of patents [33]. Erzurumlu and Pachamanova (2020) applied topic modeling to patent TI for effective technology identification and used it as an input variable. Marco and Miller (2018) analyzed the factors of patent litigation through survival time regression [34]. Marco and Miller (2018) used only QI and showed that the number of citations correlated with the occurrence of litigation. Prior studies that applied survival analysis performed patent registration prediction and patent value evaluation using various information. This paper proposes an integrated survival model to predict patent litigation hazards over time. In the proposed model, the event of interest is set as a change in litigation status. Therefore, patents for which the occurrence of litigation is unknown based on the time of data collection are considered right-censored data.

2.3. Random Survival Forest

RSF is an extension of the random forest (RF) approach proposed by Breiman (2001) and uses the ensemble tree for survival analysis [35–37]. RSF differs from RF because, when growing individual trees, information on survival time and right-censored must be included and data separation is measured according to survival difference. RSF uses bootstrap samples and randomly selected variables and is a non-parametric method [36]. The process of performing RSF is similar to that of RF and is as follows. First, create B bootstrap samples from original data. At this time, on average, 63% of the original data is used for training as in-bag data, and 37% of the bootstrap sample, which is the out-of-bag data (OOB data), is used for model verification. Next, each survival tree is grown from the generated bootstrap samples. p candidate variables are randomly selected from each node of the tree. In addition, among the p variables, the node is partitioned using the variable which maximizes the difference in patent litigation status between daughter nodes. Each tree is grown to full size until it reaches the stopping criterion set by the analyst. The final nodes created through this process are called terminal nodes. Finally, the cumulative hazard function (CHF) of each tree is calculated and the average used to obtain the ensemble CHF. h is the node corresponding to the h -th of each tree. At this time, the CHF for h is the Nelson–Aalen estimator and is as shown in Equation (1).

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \quad (1)$$

$t_{l,h}$ is the time when the l -th individual patent occurred at the h -th node. $d_{l,h}$ is the number of events that have occurred, and $Y_{l,h}$ is the number of individual patents. All individual patents included in the same node have the same CHF. A specific individual patent i is described as the d -dimensional covariate x_i . The CHF of a specific individual patent i is as shown in Equation (2) below.

$$H(t|x_i) = \hat{H}_h(t), \text{ if } x_i \in h \quad (2)$$

The CHF of the tree created in the b -th in-bag bootstrap sample is $H_b^*(t|x_i)$. The ensemble CHF of individual patent i is calculated by averaging B trees and is as shown in Equation (3).

$$H_e^*(t|x_i) = \frac{1}{B} \sum_{b=1}^B H_b^*(t|x_i) \quad (3)$$

If the individual patent i belongs to the OOB data, $I_{i,b} = 1$, otherwise $I_{i,b} = 0$. The CHF of individual patent i using OOB data is as shown in Equation (4).

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^B I_{i,b} H_b^*(t|x_i)}{\sum_{b=1}^B I_{i,b}} \quad (4)$$

In this paper, we propose a patent litigation hazard prediction model based on RSF. When applying the RSF, right-censored data and time are set according to the patent litigation status.

3. Proposed Methodology

In this paper, we propose an integrated survival model using clustering and RSF to predict patent litigation hazards. The proposed model is largely composed of three parts. First, in part 1, the data required for model construction is extracted and preprocessed while solving the label imbalance problem of litigation patents. To this end, QI is extracted from the collected patent big data, and non-litigation patents with similar propensity scores to litigation patents are sampled at a ratio of 1:5 through propensity score matching (PSM). In addition, stop word removal and lemmatization are performed using text mining techniques for title, abstract, and patent claims. In part 2, the patent is embedded in a

specific space through pre-trained word2vec. The technology fields are identified by using K-means clustering in embedded patents and technology field labels are assigned to each patent. In part 3, the extracted QI and technology label are merged to build an integrated survival model. Finally, testing data is input to the built model to evaluate its predictive performance for patent litigation hazards. Figure 1 below is a schematic diagram of the methodology proposed in this study.

Figure 1. Proposed methodology. Part 1 consists of data preprocessing for model construction and evaluation, Part 2 technology field labeling, and Part 3 is for developing the patent litigation prediction model.

3.1. Data Description and Preprocessing

This paper proposes an integrated survival model to effectively predict patent litigation hazards. Actual patent big data is collected for both training and performance evaluation of the proposed model. The patents to be collected are related to sensor semiconductors based on AI technology and are registered with the USPTO. This technology is a convergence field applicable to robotics. The application period of the collected patents is from 1998 to 2020. Out of a total of 14,198 patents, 297 have been litigated so far. Table 1 below shows information about the raw data initially collected.

Table 1. Information on collected raw data.

Technological Field	DB	Period	Status	Number of Patents (Number of Litigations)
Sensor semiconductor based on AI	USPTO	1998~2020	Registered	14,198 (297)

The litigation label imbalance problem occurs when the raw data shown in Table 1 is directly trained on the model [19,21]. This hinders the effective learning of the integrated survival model. Therefore, non-litigation patents with similar characteristics to the litiga-

tion patents are sampled at a specific rate through PSM. To this end, PSM is applied by extracting the QI shown in Table 2 from the raw data.

Table 2. Description of quantitative information.

Quantitative Information	Description	Measurable Value Information
all_IPC_count	Number of IPC codes	Technology scalability
citation_count	Number of backward citations	Technology impact
forward_citation_count	Number of forward citations	Technology impact
family_nation_count	Number of family nations	Market impact
family_doc_count	Number of family patents	Market impact
all_claim_count	Number of claims	Rights
inventor_count	Number of inventors	Sustainable development
transfer_yn	Technology transfer status (dummy)	Utility value
app_to_regi (days)	Time required from application to registration	Utility value

In Table 2, all_IPC_count is the IPC code associated with the patent. The more IPC codes assigned, the more expandable is the patent to various technologies. The citation_count and forward_citation_count indicate the number of Backward/Forward citations and the degree of citation of the patent, allowing measurement of the technology impact. Patents are effective only in the country in which the rights are obtained. Therefore, in order to secure high marketability, applications are made to many countries. The family_nation_count and family_doc_count refer to the degree to which patents have been applied for in various countries, so market influence can be measured. The all_claim_count indicates the scope of the technology protection of the patent, and it is possible to measure the rightness. This is because claims represent technical characteristics for protecting rights, so the larger the number of claims, the wider the scope of rights. transfer_yn indicates whether the applicant of the relevant patent has transferred it to another person, meaning that the transferred patent has a high utility value. app_to_regi represents the number of days from patent application to patent registration. In general, patents of excellent utility value require accelerated processing by the applicant for rapid registration. For this reason, if app_to_regi has a small value, it means that the utility value is high. Based on the extracted QI, litigation patents and non-litigation patents are matched in a 1:5 ratios and sampled. Additionally, the Wilcoxon rank sum test is used for all variables except transfer_yn, a dummy variable, to test the statistical homogeneity between the matched patents. Transfer_yn is tested using the Chi square test. Table 3 below summarizes the characteristics of the data extracted through PSM and the statistical homogeneity test results.

Table 3. Result of PSM and statistical homogeneity test.

Variables	before PSM (Mean $\pm$ SD or $n(\%)$)			after PSM (Mean $\pm$ SD or $n(\%)$)	
	Litigation ($n = 297$)	Non-Litigation ($n = 13,901$)	p -Value	Non-Litigation ($n = 1485$)	p -Value
all_IPC_count	2.48 $\pm$ 3.67	3.31 $\pm$ 3.53	<0.001	1.96 $\pm$ 1.88	0.415
citation_count	75.59 $\pm$ 117.46	40.89 $\pm$ 138.39	<0.001	76.54 $\pm$ 326.41	<0.001
forward_citation_count	77.37 $\pm$ 108.1	21.04 $\pm$ 43.03	<0.001	58.17 $\pm$ 92.94	0.002
family_nation_count	5.74 $\pm$ 5.95	2.93 $\pm$ 2.84	<0.001	5.56 $\pm$ 4.81	0.485
family_doc_count	73.29 $\pm$ 171.31	22.4 $\pm$ 138.37	<0.001	68.35 $\pm$ 240.30	<0.001
all_claim_count	35.36 $\pm$ 27.11	20.77 $\pm$ 13.33	<0.001	29.04 $\pm$ 26.39	<0.001
inventor_count	2.7 $\pm$ 1.78	3.01 $\pm$ 2.16	0.006	2.66 $\pm$ 1.81	0.863
transfer_yn					
Y	162 (54.55%)	3779 (27.19%)	<0.001	849 (57.17%)	0.441
N	135 (45.46%)	10,122 (72.82%)		636 (42.83%)	
app_to_regi (days)	1189.94 $\pm$ 675.14	1247.6 $\pm$ 691.25	0.126	1311.81 $\pm$ 680.38	0.002

When performing the statistical homogeneity test between the litigation patent and non-litigation patent groups, H_0 and H_1 mean the following:

- H_0 : There is no difference in QI between the two groups.
- H_1 : There is a difference in QI between the two groups.

As shown in Table 3, it was found that the QIs between the two groups were dissimilar before performing PSM. However, after sampling non-litigation patents through PSM, most of the QIs were found to be similar. Figure 2 below shows the propensity score of the data sampled through PSM as a plot.

Figure 2. Result of PSM plot.

As shown in Figure 2 above, the propensity score of the matched litigation patent group and non-litigation patent group are similar. The extracted 1782 patents (litigation patents = 297, non-litigation patents = 1485) are divided into training data ($n = 1426$) and testing data ($n = 356$) for training and testing the integrated survival model. In addition, TI, including the technological description of the patent, is extracted and preprocessed using the text mining technique. Preprocessing applies lemmatization to TI to extract the basic form of the words and remove numbers, punctuation marks, and stop words.

3.2. Patent Embedding and Technology Field Labeling

The patent contains the details of the developed technology in text form, so the details of the technology can be interpreted through the patent text. To identify a technology field by applying K-means clustering to the TI extracted from data, patents must be vectorized. To this end, this paper uses word2vec, which preserves context information and can effectively embed documents in specific vector spaces [38,39]. word2vec can perform effective document embedding, but requires a huge corpus volume [40]. In this paper, since the corpus volume of the patents used for model construction is very small, the embedding performance of word2vec is unlikely to be effective. To solve this problem, in this paper, we use a pre-trained word2vec, which is trained in a news corpus of large volume [41]. Table 4 below shows the characteristics of pre-trained word2vec.

Table 4. Information of pre-trained word2vec.

Model	Corpus Size (Words)	Number of Dimensions	Training Algorithm
GoogleNews-vectors-negative300	3,000,000	300	Negative sampling

The pre-trained word2vec has learnt from the Google News corpus and contains the result of embedding 3 million words into a 300-dimensional space as a real number [41]. In this paper, the word list and Document-term matrix (DTM) are derived by performing preprocessing on the patent TI. Using the word list, the matching embedded values are

called from the pre-trained word2vec, and a look-up table specific to the patent is constructed. Finally, a patent embedding matrix is derived through matrix multiplication of the DTM and the look-up table. Figure 3 below is a schematic diagram of the patent embedding process.

Figure 3. Patent embedding process.

By applying K-means clustering to the patent embedding matrix derived through the process shown in Figure 3 above, technology clusters are formed and technology fields identified [42]. When performing K-means clustering, K, which is the number of clusters, is derived through the silhouette score, which is shown in Equation (5).

$$s_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ Silhouette score} = \sum_{i=1}^n s_i \quad (5)$$

In Equation (5), i means an individual patent in the technology cluster. $a(i)$ is the average of the distances between patents in the technology cluster to which individual patent i belongs. $b(i)$ is the average of the distance between individual patent i and the patents belonging to other technology clusters. The larger the silhouette score, the more appropriate is K. Clustering is performed using the derived optimal K. Next, the technology field is defined through words with a high frequency of appearance by technology cluster. Finally, labels are assigned according to the technology field to which the patent belongs.

3.3. Patent Litigation Prediction Model

In this paper, we propose a method of predicting litigation hazards over time taking into account varied patent information and censored data. To this end, the QI extracted in part 1 and the technology field label derived in part 2 are merged and used. Moreover, censored data is set based on the litigation status up to the point of data collection. The

integrated survival model proposed in this paper uses a random survival forest based on an ensemble tree during survival analysis. Figure 4 is a schematic diagram of the construction process of the proposed integrated survival model.

Figure 4. Integrated survival model constructing process.

In Figure 4, the QI of the integrated survival model are the indicators specified in Table 2. The technology field label is given to individual patents through the patent embedding matrix and the K-means clustering performed in Part 2. Finally, the censoring status for constructing the survival model is litigation status. In addition, the censoring time of the litigation patent is the number of months from the application date to the occurrence of litigation, and for the non-litigation patent is the number of months from the application date to the data collection time. The integrated survival model proposed in this paper predicts the patent litigation hazard over time. Table 5 shows the parameters of the built integrated survival model.

Table 5. Parameters of the integrated survival model.

Components	Candidates
Splitting rule	Logrank
Number of trees	1000
Number of variables	4

The proposed integrated survival model is compared with various survival analysis models to evaluate its predictive performance for patent litigation hazard. The models used for the comparison were the Kaplan–Meier (KM), Cox proportional hazard (CoxPH), and RSF model using different splitting rules. Table 6 summarizes the details of models compared.

Table 6. Information of comparison models.

Model	Description
KM	Kaplan–Meier Estimator
Full_Cox	CoxPH model with all variables
Reduced_Cox	CoxPH model with some variables by stepwise method
RSF_brierscore	RSF using Brier score as splitting rule
RSF_logrankscore	RSF using log rank score as splitting rule

In Table 6, KM estimates the hazard function non-parametrically and is a widely used model in survival analysis. Full_Cox is the CoxPH model using all QI and technology field labels as input variables. Reduced_Cox is a CoxPH model that extracts and uses meaning-

ful variables through stepwise selection. The RSF_brierscore and RSF_logrankscore use different splitting rules from the proposed model when growing each tree. In this paper, two indicators are used to evaluate the litigation hazard prediction performance of the integrated survival model and the compared models. They are prediction error curves and the Concordance index (C-index), which are mainly used to measure the performance of the survival analysis models [35,36]. The lower the value of prediction error curves and the higher the C-index value, the better the prediction performance.

4. Experiment Results

An experiment is conducted to evaluate the predictive performance of the proposed integrated survival model for patent litigation hazard. To this end, patents related to sensor semiconductors based on AI technology are collected and samples are extracted with PSM. As a result of sampling through PSM, a total of 1782 experimental data samples were obtained, of which 1426 were used as training data and 356 used as testing data. The total number of words preprocessed to TI extracted from the training data is 5058. Using this word list, embedded values were called from pre-trained word2vec and a patent embedding matrix was constructed. K-means clustering was performed to identify technology fields. At this time, the optimal number of K technology clusters was derived through silhouette scoring as shown in Figure 5.

Figure 5. Result of silhouette score.

The optimal K is derived to be 2, as shown in Figure 4. Next, the technology fields are identified through the high frequency words associated with each cluster, and field labels are accordingly assigned to each patent. Table 7 shows information on high-ranking words by cluster and the corresponding defined technology fields.

Table 7. Information of technology fields.

Cluster (Technology Label)	Top Frequency Word List	Technology Field Definition
Cluster 1 (Tech 1)	signal, control, detect, plural	Object recognition technology
Cluster 2 (Tech 2)	Communic, receiv, network	Signal communication technology

Figure 6 shows patent litigation hazard as CHF for each model over time using the training data. As shown in Figure 6, for all models CHF increases steadily over time after a patent is filed. This is because as time goes by, patents are released to more people and similar technologies are developed, so litigation is more likely to occur. In particular, it was found that the litigation hazard began to occur 150 months after the patent application for all models, and increased rapidly after 200 months.

Figure 6. Cumulative hazard function by model.

The variable importance (VIMP) and hazard ratio (HR) according to the models are shown in the following tables to understand the influence of the input variables on the occurrence of patent litigation. Table 8 shows the VIMP of the integrated survival model, RSF_brierscore, and RSF_logrankscore built using training data. VIMP represents the degree to which each variable in the model affects the patent litigation hazard.

Table 8. Variable importance with RSF models.

Variable	Integrated Survival Model (Proposed Model)	RSF_brierscore	RSF_logrankscore
all_IPC_count	0.0240	0.0147	0.0110
citation_count	0.0317	0.0301	0.0174
forward_citation_count	0.0215	0.0175	0.0183
family_nation_count	0.0096	0.0102	0.0102
family_doc_count	0.0142	0.0172	0.0178
all_claim_count	0.0180	0.0208	0.0185
inventor_count	0.0002	0.0008	0.0019
transfer_yn	0.0101	0.0081	0.0066
app_to_regi	0.0113	0.0150	0.0192
Technology field label	0.0007	0.0006	0.0012

In Table 8, the VIMP of the RSF models differed, but in general, citation_count, app_to_regi, all_claim_count, and all_IPC_count were derived to be high influence variables. Table 9 summarizes the HR for each variable with the CoxPH models. A higher HR value indicates more influence on patent litigation hazard.

As shown in Table 9, in the CoxPH models, transfer_yn, all_IPC_count, and the technology field label were derived as variables that affect the patent litigation hazard. To evaluate the learning performance of the proposed model and the models used for comparison, a 10-fold cross validation was performed, and the prediction error curves and C-index derived. Figure 7 shows the performance of the models after training with the training data.

Table 9. Hazard ratio by variable with CoxPH models.

Variables	Full_Cox HR (95% CI)	Reduced_Cox HR (95% CI)
all_IPC_count	1.18 (1.14–1.2)	1.18 (1.14–1.2)
citation_count	1.00 (1.00–1.0)	1.00 (1.00–1.0)
forward_citation_count	1.00 (1.00–1.0)	1.00 (1.00–1.0)
family_nation_count	1.03 (1.01–1.0)	1.03 (1.02–1.0)
family_doc_count	1.00 (1.00–1.0)	-
all_claim_count	1.00 (1.00–1.0)	-
inventor_count	0.97 (0.94–1.0)	0.97 (0.94–1.0)
transfer_yn		
Y	1.37 (1.22–1.5)	1.36 (1.21–1.5)
N	Reference	Reference
app_to_regi	1.00 (1.00–1.0)	1.00 (1.00–1.0)
Technology field label		
Tech 1	Reference	Reference
Tech 2	1.18 (1.04–1.3)	1.18 (1.04–1.3)

Figure 7. Result of training models. (a) Prediction error curves for trained models; (b) C-index for trained models.

The prediction error curves in Figure 7 show that the values of all models increase over time. Of them, KM has the largest prediction error values while the proposed model has the lowest prediction error values. In the case of C-index, the value of KM is almost unchanged from 0.5, and the CoxPH models have values ranging between 0.9 and 0.7. In addition, the RSF models mostly maintain high C-index values.

In this paper, we use the previously put aside 356 test data samples to determine whether the proposed model performs well at predicting litigation hazards for the new patent input data. The process of model performance evaluation using the test data follows. First, the patent embedding matrix with test data is constructed as in Part 1. Next, by measuring the centroid points of the technology fields derived from the training data by Euclidean distance, a technology field label is given to the patents of the test data. Finally, we calculate the prediction error and C-index of the test data with the trained models, and evaluate prediction performance. Table 10 compares the mean and standard deviation (SD) of the prediction performance indicators derived using the test data.

Table 10 shows that the RSF models showed better prediction performance than the KM and CoxPH models. Among them, the mean prediction error and SD of the proposed model are 0.11 ± 0.07 , and the C-index mean and SD are 0.81 ± 0.14 , indicating good prediction performance. Moreover, for both indicators, KM showed low prediction performance, and the CoxPH models had the same prediction error, mean and SD of the C-index. As shown in Table 10, even with testing data, that is, when new data is input, it can be seen that the proposed model performs well for predicting the patent litigation hazard.

Table 10. Comparison of prediction result.

Model	Prediction Error (Mean $\pm$ SD)	C-Index (Mean $\pm$ SD)
KM	0.14 $\pm$ 0.09	0.5 $\pm$ 0.001
Full_Cox	0.13 $\pm$ 0.08	0.72 $\pm$ 0.16
Reduced_Cox	0.13 $\pm$ 0.08	0.72 $\pm$ 0.16
Proposed model	0.11 $\pm$ 0.07	0.81 $\pm$ 0.14
RSF_brierscore	0.12 $\pm$ 0.07	0.8 $\pm$ 0.14
RSF_logrankscore	0.12 $\pm$ 0.08	0.81 $\pm$ 0.15

5. Discussion

A patent gives a company exclusive rights to the developed technology, but, like a double-edged sword, it can become the target of infringement litigation by another company. Patents contain detailed information about the developed technology. However, the characteristics of the technology described in the patent may infringe the scope of the rights of other companies, so litigation may occur. Patent litigation adversely affects the sale of products and services of a company, hindering the sustainability of business activities. For this reason, many studies have proposed methods to prevent litigation. However, the existing studies do not reflect patent litigation hazard and censored data over time.

In this study, an integrated survival model was proposed to predict the patent litigation hazard over time. The proposed model can reflect both patent QI and TI, and experimentally demonstrated better prediction performance than the compared survival analysis models. In survival analysis, as in previous studies on quantitative methods, the tree-based model showed better predictive performance than the regression model. In addition, the number of forward/backward citations, which were mainly used as variables of models in previous studies, was found to have a great influence on the occurrence of patent litigation. For this reason, companies need to carefully manage patents with a large number of citations for litigation prevention and sustainable management. The proposed model can be used in various fields of fourth industrial revolution technology, including robotics. However, there are some limitations to this study. First, the type of patent litigation was not considered. Actually, patent litigation types include direct infringement, indirect infringement, and others. Therefore, considering the appropriate method of extracting and processing patent information according to patent litigation type is necessary. Second, in the proposed model, the TI's preprocessing method is done only in English. Patents are written and applied in various languages in many countries. Thus, in order to apply the proposed model to various countries, the TI preprocess should consider a variety of languages.

6. Conclusions

Patent litigation occurs when a company's patents, products, or operations infringe on the scope of other companies' patent rights. When patent litigation occurs, much time and expense is needed to resolve it. Therefore, companies implement various strategies to preempt patent litigation. Among them, quantitative litigation prediction methods using statistics and machine learning are efficient and can produce objective results. For this reason, many studies have been conducted to predict patent litigation using quantitative methods. Previous studies have constructed prediction models using varied patent information. However, they do not take into account the characteristics of litigation hazards changing over time.

In this paper, we proposed the integrated survival model to predict patent litigation hazard over time. The proposed model considers the QI and TI of the patent, and censored data based on RSF can also be used. Moreover, when using TI to identify technology fields, a pre-trained word2vec was used to complement the small corpus size. Finally, patent data was collected and used to examine the industrial applicability of the proposed model.

The patent is actually registered with the USPTO and is related to sensor semiconductors based on AI. Through the experimental results, the proposed model demonstrated better prediction performance than other survival analysis models. The proposed model could be used by potential defendants to protect their patents.

The limitations of this study discussed in Section 5 need to be resolved through future studies. First, various types of patent litigation should be collected and their characteristics identified. Indicators that can reflect the information of each litigation type should be appropriately selected. These indicators need not be inherent to the patents themselves, but could also use other variables, such as those related to the patent owner's finances and competition. Finally, a deep learning-based natural language processing method should be adapted to process the patent TI of different countries.

Author Contributions: Y.K., S.P., J.L. and J.K. conceived and designed the experiments; D.J. and J.L. participated in the data collection to confirm the validity of this study; Y.K. wrote the paper and performed all of the research steps. In addition, Y.K., S.P., J.L. and J.K. cooperated with each other in revising the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MOTIE (Ministry of Trade, Industry, and Energy) in Korea, under the Fostering Global Talents for Innovative Growth Program (P0008749) supervised by the Korea Institute for Advancement of Technology (KIAT).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to restrictions [contact: youngho0928@korea.ac.kr].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eisenberg, R.S. Patents and the progress of science: Exclusive rights and experimental use. *Univ. Chic. Law Rev.* **1989**, *56*, 1017–1086. [CrossRef]
2. Mogee, M.E. Using patent data for technology analysis and planning. *Res. Technol. Manag.* **2016**, *34*, 43–49. [CrossRef]
3. Moro-Visconti, R. *The Valuation of Digital Intangibles: Technology, Marketing and Internet*; Palgrave Macmillan: London, UK, 2020; pp. 127–153, ISBN 978-3-030-36917-0.
4. World Intellectual Property Organization. *World Intellectual Property Indicators*; WIPO: Geneva, GN, Switzerland, 2019; pp. 1–223.
5. Casey, K.R. Alternate dispute resolution and patent law. *Fed. Cir. B. J.* **1993**, *3*, 1–14.
6. Llobet, G. Patent litigation when innovation is cumulative. *Int. J. Ind. Organ.* **2003**, *21*, 1135–1157. [CrossRef]
7. Bessen, J.; Meurer, M.J. Lessons for patent policy from empirical research on patent litigation. *Lewis Clark L. Rev.* **2005**, *9*, 1–28.
8. Chien, C.V. Startups and patent trolls. *Stanf. Technol. Law Rev.* **2012**, *17*, 461–506. [CrossRef]
9. Graevenitz, G.V.; Wagner, S.; Harhoff, D. Incidence and growth of patent thickets: The impact of technological opportunities and complexity. *J. Ind. Econ.* **2013**, *61*, 521–563. [CrossRef]
10. Lee, J. Essays on Market for Technology and Patent Strategy: Licensing, Patent Transaction, and Patent Litigation. Ph.D. Thesis, Korea Advanced Institute of Science and Technology, Daejeon, Korea, 2018.
11. Yang, X.; Yu, X. Preventing Patent risks in artificial intelligence industry for sustainable development: A multi-level network analysis. *Sustainability* **2020**, *12*, 8667. [CrossRef]
12. Lai, Y.; Che, H. Modeling patent legal value by extension neural network. *Expert Syst. Appl.* **2009**, *36*, 10520–10528. [CrossRef]
13. Chien, C.V. Predicting patent litigation. *Tex. Law Rev.* **2011**, *90*, 283–329.
14. Kim, A.; Partee, N.W.; Reynolds, T.; Santamaria, M.A.; Advisor, F.; Beling, P.; Ferron, D.V.; Winer, D.S.; Woolley, D.F. Patent litigation risk-scoring model. Available online: <https://www.semanticscholar.org/paper/PATENT-LITIGATION-RISK-SCORING-MODEL-Kim-Partee/e634d98fc4759aff3c3409c2315a0297ec71e7bf> (accessed on 12 November 2020).
15. Lanjouw, J.O.; Schankerman, M. Characteristics of patent litigation: A window on competition. *RAND J. Econ.* **2001**, *32*, 129–151. [CrossRef]
16. Bessen, J.; Meurer, M.J. The patent litigation explosion. *Loy. U. Chi. L. J.* **2013**, *45*, 401–440. [CrossRef]
17. Lee, J.; Kang, J.; Jun, S.; Lim, H.; Jang, D.; Park, S. Ensemble modeling for sustainable technology transfer. *Sustainability* **2018**, *10*, 2278. [CrossRef]
18. Trappey, C.V.; Trappey, A.J.C.; Wu, C. Clustering patents using non-exhaustive overlaps. *J. Syst. Sci. Syst. Eng.* **2010**, *19*, 162–181. [CrossRef]
19. Kim, C. A Study on Formulating Patent Litigation Forecasting Model and Method to Process Litigation Evaluation Factors. Ph.D. Thesis, Korea Advanced Institute of Science and Technology, Daejeon, Korea, 2014.

20. Cowart, T.W.; Lirely, R.; Avery, S. Two Methodologies for Predicting Patent Litigation Outcomes: Logistic Regression Versus Classification Trees. *Am. Bus. Law J.* **2014**, *51*, 843–877. [CrossRef]
21. Wongchaisuwat, P.; Klabjan, D.; McGinnis, J.O. Predicting litigation likelihood and time to litigation for patents. *arXiv* **2016**, arXiv:1603.07394.
22. Kartsonaki, C. Survival analysis. *Diagn. Histopathol.* **2016**, *22*, 263–270. [CrossRef]
23. Rupert, G.M. Survival analysis. In *Introduction to Survival Concepts*; John Wiley & Sons: New York, NY, USA, 1998; pp. 1–9.
24. Machin, D.; Cheung, Y.B.; Parmar, M. Survival analysis: A practical approach. In *Introduction and Review of Statistical Concepts*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006; pp. 1–22.
25. Guerzoni, M.; Nava, C.R.; Nuccio, M. The survival of start-ups in time of crisis. A machine learning approach to measure innovation. *arXiv* **2019**, arXiv:1911.01073.
26. Leung, K.; Elashoff, R.M.; Afifi, A.A. Censoring issues in survival analysis. *Annu. Rev. Public Health* **1997**, *18*, 83–104. [CrossRef]
27. Wagner, S.; Cockburn, I. Patents and the survival of Internet-related IPOs. *Res. Policy* **2010**, *39*, 214–228. [CrossRef]
28. Zeebroeck, N.V. Patents Only Live Twice: A Patent Survival Analysis in Europe. Available online: https://www.researchgate.net/publication/24131120_Patents_Only_Live_Twice_A_Patent_Survival_Analysis_in_Europe (accessed on 12 November 2020).
29. Xie, Y.; Giles, D.E. A survival analysis of the approval of us patent applications. *Appl. Econ.* **2011**, *43*, 1375–1384. [CrossRef]
30. Nikzad, R. Survival analysis of patents in canada. *J. World Intellect. Prop.* **2011**, *14*, 368–382. [CrossRef]
31. Nakata, Y.; Zhang, X. A survival analysis of patent examination requests by japanese electrical and electronic manufacturers. *Econ. Innov. New Technol.* **2012**, *21*, 31–54. [CrossRef]
32. Han, E.J.; Sohn, S.Y. Patent valuation based on text mining and survival analysis. *J. Technol. Transf.* **2015**, *40*, 821–839. [CrossRef]
33. Erzurumlu, S.S.; Pachamano, D. Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. *Technol. Forecast. Soc. Chang.* **2020**, *156*, 120041. [CrossRef]
34. Marco, A.C.; Miller, R.D. *Patent Value and Uncertain Property Rights: Implications from Patent Litigation*; Working papers Hoover IP2: Stanford, CA, USA, 2018; p. 18008.
35. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [CrossRef]
36. Weathers, B.; Cutler, D.R. Comparison of Survival Curves between Cox Proportional Hazards, Random Survival Forest, and Conditional Inference Forests in Survival Analysis. Available online: <https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1936&context=gradreports> (accessed on 12 November 2020).
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
39. Rong, X. Word2vec parameter learning explained. *arXiv* **2016**, arXiv:1411.2738.
40. Rezaeinia, S.M.; Rahmani, R.; Ghodsi, A.; Veisi, H. Sentiment analysis based on improved pre-trained word embeddings. *Expert Syst. Appl.* **2019**, *117*, 139–147. [CrossRef]
41. Google LLC. Word2vec. Available online: <https://code.google.com/archive/p/word2vec/> (accessed on 12 November 2020).
42. Sharma, A. A survey on different text clustering techniques for patent analysis. *Int. J. Eng. Res. Technol.* **2012**, *1*, 1–4.